COMPLEX NETWORKS 2019

THE 8TH INTERNATIONAL CONFERENCE ON COMPLEX NETWORKS AND THEIR APPLICATIONS

December 10 - 12, 2019
Lisbon, Portugal

BOOK OF ABSTRACT

COMPLEX NETWORKS 2019
The 8th International Conference on Complex Networks & Their Applications
December 10 - 12, 2019 Lisbon, Portugal
Published by the International Conference on Complex Networks & Their Applications.

Editors:
Hocine Cherifi
University of Burgundy, France

José Fernando Mendes
University of Aveiro, Portugal

Luis Mateus Rocha
Indiana University, USA

Sabrina Gaito,
University of Milan, Italy

Esteban Moro
Universidad Carlos III, Spain

Joana Gonçalves-Sá
Universidade Nova de Lisboa, Portugal

Francisco Santos
University of Lisbon, Portugal

COMPLEX NETWORKS 2019
e-mail: hocine.cherifi@u-bourgogne.fr

ISBN: 978-2-9557050-3-2

Preface

We are proud to present the Book of Abstracts for the 8th International Conference on Complex Networks & their Applications: COMPLEX NETWORKS 2019 Since 2012 the event has been held around the world on a yearly Basis. After Sorrento (Italy), Kyoto (Japan), Marrakech (Morocco), Bangkok (Thailand), Milan (Italy), Lyon (France), Cambridge (UK) the eighth edition is hosted by the Gulbenkian Science Institute, in Lisbon from December 10 to December 12, 2019. The originality of the conference lies in the strongly interdisciplinary nature of the topics covered. Indeed, complexity and network science are multidisciplinary fields that mobilize intellectual resources in virtually all-scientific communities. Nowadays, all disciplines (physics, biology, social sciences, economics, computer science, meteorology, etc.) are faced with a massive influx of data and an explosion of information to manage. Through the data and their interactions, network science aims at understanding these complex systems increasingly large. COMPLEX NETWORKS is very focused at being an interdisciplinary event. However, this is linked with willingness to the requirements that the quality of the contributions must be among the best work in each of the scientific fields covered. In order to guarantee the excellence and reputation of this event, for its eighth edition COMPLEX NETWORKS has brought together in its scientific committee more than 400 leading international experts from all over the world. Year after year the event has increased its international influence. The 470 contributions that we received this year, from more than 50 countries around the world have been peer reviewed by at least 3 independent reviewers. This publication gathers the 190 extended abstracts accepted for presentation together with abstracts of six keynote speeches and two invited tutorials.

Each edition of the conference represents a challenge that cannot be successfully achieved without the deep involvement of plenty of people, institutions and sponsors. We would like to thank all of them. We record our thanks to our fellow members of the Organizing committee for their huge efforts for the success of the conference. The program committee members for their engagement in promoting the event and refereeing submissions as well as the local committee members for their great commitment over the past months. We are also indebted to our sponsors, in particular Tribe Communication for designing the visual identity of the Conference. We are equally grateful to all the institutions that have helped us, in particular, the Calouste Gulbenkian Foundation for hosting this event. We also wish to express our appreciation to all participants and presenters. On a final note, we would like to express our deep sense of appreciation to our keynote and tutorial speakers.

Hocine Cherifi
Sabrina Gaito
Joana Gonçalves-Sá
José Fernando Mendes
Esteban Moro
Luis Mateus Rocha
Francisco C. Santos

Table of Contents

Tutorials

Invited Speakers

I Biological Networks

II Community Structure

III Diffusion and Epidemics

IV Dynamics on/of Networks

V Human Behaviour

VI Link Analysis and Ranking

VII Machine Learning and Networks

VIII Mobility

IX Multilayer Networks

X Network Analysis and Measure

XI Network Geometry

XII Network Models

XIII Network Neuroscience

XIV Networks in Finance and Economics

XV Political Networks

XVI Quantifying Success

XVII Resilience and Control

XVIII Social Networks

XIX Synchronization, Resilience and Control

XX Urban Networks

Tutorials

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Mapping networks in latent geometry: models and applications

Maria Ángeles Serrano

Universitat de Barcelona, Spain

Complex networks talk a common language, regardless of their origin, and are imprinted with universal features. Many of these features are well explained by the S1/H2 family of hidden metric space network models, where nodes are placed at specific coordinates in an underlying geometry, which led to the discovery that the effective geometry of many real networks is hyperbolic. Hyperbolicity emerges as a result of the combination of heterogeneous popularity and Euclidean similarity into an effective distance between nodes, such that more popular and similar nodes have more chance to interact. The geometric approach allows the production of truly cartographic maps of real networks in hyperbolic space that can be obtained using different techniques. Recently, we have introduced Mercator, an embedding tool that mixes machine learning and maximum likelihood approaches to perform dimensional reduction giving the coordinates of the nodes in the underlying hyperbolic disk with the best matching between the observed network topology and the underlying S1/H2 geometric model. The maps are not only visually appealing, but also meaningful and enable efficient navigation, the detection of communities of similar nodes, and a geometric renormalization group that unravels the multiple length scales coexisting in the structure of complex networks, strongly intertwined due to the small world property. The application of geometric renormalization to real networks unfolds them into a multilayer shell that shows scale invariance, meaning that the same principles are ruling the formation of network connections at different length scales. Interestingly, this self-similarity may have its origin in an evolutionary drive. Beyond its explanatory power, practical applications of the geometric renormalization technique include multiscale navigation and the production of downscaled or upscaled network replicas, among many other.

M. Ángeles Serrano obtained her Ph.D. in Physics at the Universitat de Barcelona in 1999 with a thesis about gravitational wave detection. One year later, she also received her Masters in Mathematics for Finance from the CRM-Universitat Autònoma de Barcelona. After four years in the private sector as IT consultant and mutual fund manager, she returned to academia in 2004 to work in the field of complex networks. She completed her postdoctoral research at Indiana University (USA), the École Polytechnique Fédérale de Lausanne (Switzerland) and IFISC Institute (Spain). She came back to Barcelona in 2009, when she was awarded a Ramón y Cajal Fellowship at UB. In February 2009, she obtained the Outstanding Referee award

of the American Physical Society. She is a Founder Member of Complexitat, the Catalan Network for the study of Complex Systems, and a Promoter Member of UBICS, the Universitat de Barcelona Institute of Complex Systems. M. Ángeles Serrano is ICREA Research Professor at the Universitat de Barcelona from October 2015.

Wikimedia Public (Research) Resources

Diego Saez-Trumper

Wikimedia Foundation

The Wikimedia Foundation's mission is to disseminate open knowledge effectively and globally. In keeping with this mission, the Wikimedia Foundation support research in areas that benefit the Wikimedia community. We aim to make any work with our support openly available to the public. At the same time that we do a minimalist user data collection, all the material (text and multimedia) available in our projects is public and reusable by everybody. Moreover, all the article versions and interactions among users are also public, and we offer a set of tools for accessing such data. In this tutorial we are going to give an overview on all the data sources, and a detailed explanation of how to interact with this content, including data and tools such as the Wikipedia Dumps, Quarry (SQL Replicas), Pageviews, PAWS (Jupyter Public Notebooks), Wikimedia Commons (multimedia content) and WikiData.

Diego Sáez-Trumper is a Research Scientist at Wikimedia Foundation. Before, he was a post-doctoral researcher at Yahoo! Labs (Barcelona), Senior Research Scientist at Eurecat , Data Scientist at NTENT, and part time lecturer at UPF. He holds a diploma on Acoustic Engineering (Universidad Austral de Chile, 2006) and obtained his Phd in Information Technology from Universitat Pompeu Fabra (2013) under the supervision of Dr. Ricardo Baeza-Yates. During his PhD he interned at Qatar Computing Research Institute (2013), University of Cambridge (2012) and Universidade Federal de Minas Gerais (2011). His research interests include: Diffusion of information, innovation, and influence in online social networks; User modeling; Free knowledge; Relationship between social and mainstream media; Algorithms on graphs; and privacy issues.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Invited Speakers

Reflections of social networks

Lada Adamic

Facebook

In this talk I will describe two studies based on friendship ties on Facebook. In the first, aggregate county-to-county ties in the United States tell of geographical distance but also characteristics of the counties and past migrations between them. In the second, we show how college social networks take shape, a process influenced by the type of college and the seasonality of academic life

Lada Adamic leads the Computational Social Science Team at Facebook. Prior to joining Facebook she was an associate professor at the University of Michigan's School of Information and Center for the Study of Complex Systems. Her research interests center on information dynamics in networks.

Network-based dynamic modeling of biological systems: toward understanding and control

Reka Albert

Penn State University

My group is using network science to understand the emergent properties of biological systems. As an example, we think of cell types as attractors of a dynamic system of interacting (macro)molecules, and we aim to find the network patterns that determine these attractors. We collaborate with wet-bench biologists to develop and validate predictive dynamic models of specific systems. We then use the specific knowledge gained to draw general conclusions that connect a network's structure and dynamics. An example of such a general connection is our identification of stable motifs, self-sustaining cyclic structures in the network that determine a trap subspace of the system's state space, or equivalently determine points of no return in the dynamics of the system. We have shown that control of stable motifs can guide the system into a desired attractor. Such attractor control can form the foundation of therapeutic strategies on a wide application domain. I will illustrate such applications in our model of a cell fate change that represents the first step toward cancer metastasis. Several model-predicted therapeutic interventions to block this cell fate change were validated experimentally.

Prof. Réka Albert received her Ph.D. in Physics from the University of Notre Dame (2001), working with Prof. Albert-László Barabási, then did postdoctoral research in mathematical biology at the University of Minnesota, working with Prof. Hans G. Othmer. She joined Penn State in 2003, where she currently is a Distinguished Professor of Physics with adjunct appointments in the Department of Biology and the Huck Institute of the Life Sciences. Prof. Albert is a network scientist who works on predictive modeling of biological regulatory networks at multiple levels of organization. Dr. Albert's pioneering publications on the structural heterogeneities of complex networks had a large impact on the field, reflected in their identification as "Fast breaking paper" and "High impact paper". Prof. Albert is a fellow of the American Physical Society and of the Network Science Society and an external member of the Hungarian Academy of Sciences. She was a recipient of an NSF Career Award (2007), the Maria Goeppert-Mayer award (2011), and the Distinguished Graduate Alumna Award of the University of Notre Dame (2016). Her service to the profession includes serving on the editorial board of the Biophysical Journal, Bulletin of Mathematical Biology, npj Systems Biology and Applications, and as peer reviewer for more than 35 journals.

The keynote is sponsored by Applied Network Science

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

On a Positional Approach to Network Science

Urlik Brandes

ETH Zürich

This presentation is about network science methodology. By viewing it as a data science rather than, say, a collection of methods or a unifying theory, we create opportunities for more rigorous research, both mathematically and empirically. Pivotal to the adaptation of methods to general, multivariate and temporal, situations is the notion of network position, which summarizes the relationships of a node with the rest of the network. I will give examples showcasing how the analysis of centralities, roles, and communities can benefit from a positional perspective.

Ulrik Brandes is a professor of social networks at ETH Zurich since 2017. With a background in algorithmics, his main interests are in network analysis and visualization, with application to social networks in particular. He is a co-author of the visone software for network analysis and of the GraphML data format. Deutsche Forschungsgemeinschaft (DFG) awarded him a Reinhart Koselleck-Project on Social Network Algorithmics, in which he took a shot at improving the methodological foundations of network science, and he was a principal investigator in the ERC Synergy Project NEXUS 1492 where he worked on reconstructing archaeological networks from fragmented and heterogeneous observations. Brandes received a Diploma degree from RWTH Aachen in 1994and a PhD from the University of Konstanz in 1999, both in computer science. After postdoctoral research visits to Brown University and the University of Sydney, he completed his habilitation in 2002 and became associate professor at the University of Passau the same year. From 2003-2017 he was full professor of algorithmics at the University of Konstanz. He is a member of the board of directors of the International Network for Social Network Analysis (INSNA) since 2008, and was a member of the Graph Drawing Steering Committee 2007-2014. He acts as the coordinating editor of Network Science and as an associate editor of Social Networks, and he is an editorial board member of the Journal of Mathematical Sociology as well as the Journal of Graph Algorithms and Applications.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Temporal networks: past, present, future

Jari Saramäki

Aalto University, Finland

The key strength of network science has been its ability to strip away unnecessary details, making it easier to grasp the inner workings of systems that are large and complex. At the same time, however, entire subfields have emerged that build on adding back some of this detail: weighted networks, multilayer networks, and temporal networks, the latter being the topic of this talk. I will provide an overview of what temporal networks are and what the temporal networks framework can do, and discuss when the temporal-network treatment is useful and when not. I will discuss some key findings and methods, using time-stamped social interactions as an example case, and finally, try to sketch some future directions for temporal-network research.

Jari Saramäki is a full professor and vice head at the Department of Computer Science, Aalto University, Finland. He received his PhD in applied physics in 1998, studying quantum crystals at milliKelvin temperatures. After some career twists and turns involving technology companies and what we would nowadays call data science, he returned to academia in 2003 to study complex networks, a new and rapidly expanding field at that time. Jari Saramäki is probably best known for his work on social and temporal networks, but his broad range of research interests has included topics from ant supercolonies to the human immune system.

How to eliminate systemic risk from financial multi-layer networks

Stefan Thurner

Medical University of Vienna, Austria

Given the detailed network structure of financial obligations in financial markets one can compute not only compute the systemic risk contribution of the individual financial players, but also it becomes possible to estimate the contribution of systemic risk of every single financial transaction. This in turn allows us to design incentive schemes for market participants to become systemic risk sensitive, by preferring systemically unrisky transactions. We show that such schemes lead to a restructuring of financial exposure networks in ways that suppress the possibility of cascading failure and thereby drastically reduces systemic risk. We discuss ways to compute optimal financial networks that can be used to benchmark and monitor actual financial networks.

Stefan is full professor for Science of Complex Systems at the Medical University of Vienna. He is the president of the Complexity Science Hub Vienna, external professor at the Santa Fe Institute, and a senior researcher at IIASA. Stefan obtained a PhD in theoretical physics from the Technical University of Vienna and a PhD in economics from the University of Vienna. Stefan started his career in theoretical particle physics and gradually shifted his focus to the understanding of complex adaptive systems. He published about 200 articles in physics, applied mathematics, network theory, evolutionary dynamics, life sciences, economics and finance, and lately in social sciences. He holds two patents. His work has been covered by international media such as the New York Times, BBC world, Nature, New Scientist, Physics World, and is featured in more than 400 newspaper, radio and television reports. Stefan was elected Austrian "scientist of the year" in 2018.

Machine learning for Graphs based on Kernels

Michalis Vazirgiannis

Ecole Polytechnique, France

Graph kernels have attracted a lot of attention during the last decade, and have evolved into a rapidly developing branch of learning on structured data. During the past 20 years, the considerable research activity that occurred in the field resulted in the development of dozens of graph kernels, each focusing on specific structural properties of graphs. Graph kernels have proven successful in a wide range of domains, ranging from social networks to bioinformatics. The goal of this presentation is to provide a unifying view of the literature on graph kernels. In particular, we present a comprehensive overview of a wide range of graph kernels. Furthermore, we perform an experimental evaluation of several of those kernels on publicly available datasets, and provide a comparative study. Finally, we discuss key applications of graph kernels, and outline some challenges that remain to be addressed. The experimental comparison was based on an open source python library (Grakel) we designed implementing all the known so far graph kernels.

Dr. Vazirgiannis is a Professor at LIX, Ecole Polytechnique in France. He has conducted research in Frauenhofer and Max Planck-MPI (Germany), in INRIA/FUTURS (Paris). He has been a teaching in AUEB (Greece), Ecole Polytechnique, Telecom-Paristech, ENS (France), Tsinghua, Jiaotong Shanghai (China) and in Deusto University (Spain). His current research interests are on deep and machine learning for Graph analysis (including community detection, graph classification, clustering and embeddings, influence maximization), Text mining including Graph of Words, deep learning for word embeddings with applications to web advertising and marketing, event detection and summarization. He has active cooperation with industrial partners in the area of data analytics and machine learning for large scale data repositories in different application domains. He has supervised twenty completed PhD theses. He has published three books and more than a 200 papers in international refereed journals and conferences and received best paper awards in ACM CIKM2013 and IJCAI2018. He has organized large scale conferences in the area of Data Mining and Machine Learning (such as ECML/PKDD) while he participates in the senior PC of AI and ML conferences – such as AAAI and IJCAI. He has received the ERCIM and the Marie Curie EU fellowships, the Rhino-Bird International Academic Expert Award by Tencent and between 2015 and 2018 he lead the AXA Data Science chair.

The keynote is sponsored by Frontiers in Big Data

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Part I

Biological Networks

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Network models of fracture in materials with hierarchical microstructure

Nosaibeh Esfandiary[1], Paolo Moretti[1], and Michael Zaiser[1,2]

[1] Dept. of Materials Science, WW8-Materials Simulation, FAU Universitt, Erlangen-Nrnberg, Dr.-Mack-Strae 77, 90762 Frth, Germany
nosaibeh.esfandiary@fau.de,
WWW home page: http://matsim.techfak.uni-erlangen.de
[2] School of Mechanics and Engineering, Southwest Jiaotong University Chengdu 610031, China

1 Introduction

Hierarchical materials are characterized by modules that repeat several times on different length scales in a self-similar fashion. Biological materials provide examples of heirarchical systems. Collagen protein, for instance, exhibits a hierarchical fiber organization in different scales from Angstrom until centimeter, comprises molecules, microfibrils, fibers, and fiber bundles [1]. This structure provides some properties like enhanced fracture toughness, which isolated collagen molecules can not show. Some authors [2] have suggested that hierarchical structures may delay or prevent the nucleation and spreading of critical flaws which control failure of non-hierarchical heterogeneous materials [3, 4].
Hierarchical structures play also a key role in adhesion, as it is evident in the case of the gecko. The peculiarity of this reptile is its ability to walk on ceilings and vertical walls, despite its comparably high weight. This is due to the particular fractal structure of the gecko toes, whose extremities are composed by hundreds of thousands of 100 micrometer long fibers, named setae, each of them branching in hundreds of fibrils, or spatulae, in the scale of nanometers. This structure optimizes the ability of the gecko to use van der Waals forces to adhere to a surface, even if it is rough, to detach easily and to resist flaws [5, 6].
In this work we use network models to study properties of hierarchical structures [7] and to explore how hierarchical system affects the precursor activity in the run-up to failure and ultimately changes the mode of failure, we formulate for the first time hierarchical generalizations of the well-known random fuse network (RFN) [8, 9]. At the same time we emphasize that RFN models represent a scalar caricature of tensorial elasticity. RFN models can describe fracture of materials only in exceptional cases [10]. We consider both 2- and 3-dimensional network models of hierarchical materials and we show that both bulk fracture and interface adhesion and detachment of such systems are characterized by a novel failure mode, in which crack growth is hindered at all scales.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

In our work, we generalize fuse network models in different variants to investigate the impact of hierarchical organization on failure modes and highlight differences between hierarchical and non-hierarchical materials. We consider variants of deterministic and stochastic fuse networks, both of the hierarchical (fractal) and of the non-hierarchical type. While non-hierarchical systems are characterized by structural gaps (or voids) with a well defined mean size and a short-tailed distribution, the hierarchical ones naturally display heavy-tailed power-law gap size distribution, resulting in the ability of such systems to confine crack growth.
Our numerical results confirm this picture. We use the standard Random Fuse Model to investigate the elastic response of our systems and their failure behavior under load. Fig. 1 shows the crack profiles in non-hierarchical random reference fuse network (R-RFN) and deterministic hierarchical fuse network (D-HFN), in their simpler 2-dimensional variants. The non-hierarchical systems produce the typical self-affine crack profile, as studied in the literature on RFN models[11], and pointing to fracture as a critical phenomenon: the failure point is effectively a critical point, displaying scale-invariant behavior. The hierarchical systems, instead, do not fail by growing a single large crack at the critical point, they rather accumulate micro-cracks all along the subcritical regime, resulting in highly deflected crack profile at failure. Deflections are power-law distributed in size and point to a fracture scenario that does not change when reaching the failure point. To understand better this fact, we study the distribution of avalanche sizes in the subcritical and critical regimes, which is defined as the number of links that fail without any further increase in the applied load. Once again, the non-hierarchical systems exhibit a standard phase transition behavior, with avalanche size distributions becoming power-laws at failure. In the hierarchical case, instead, no difference arises between the behavior before failure and at failure: avalanche sizes are power laws at every loading stage, with non-universal exponents that depend on the proximity of the failure point. Therefore in hierarchical systems there is no qualitative difference in the mechanical response at the failure point and before it, no precursory activity and no critical crack growth.
Our results carry over to our three dimensional models of adhesion and interface failure of hierarchical materials in contact with heterogeneous substrates, motivated by the case study of the gecko pad. In this case too, detachment of the hierarchical system is not characterized by a single catastrophic event in which the spatial symmetry of micro-crack layouts is broken in favor of a single critical crack. Micro cracks remain localized instead. This ability to confine damage results in higher fracture toughness, making the hierarchical system more effective in adhering to a heterogeneous substrates with quenched disorder.

Summary. We study numerically fracture and failure in network models of bio-inspired hierarchical materials, using the Random Fuse Model to simulate mechanical loading and breaking. We find out that unlike non-hierarchical systems, in which failure occurs as a critical phenomenon, with scale invariant behavior at a critical tipping point, in our hierarchical systems no tipping point is found and breaking processes advance by dam-

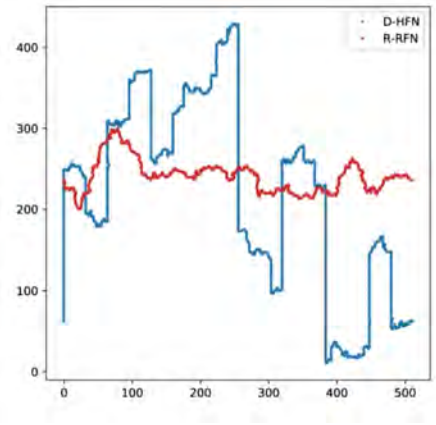


Fig. 1. Crack shape in a hierarchical (D-HFN) and a non hierarchical (R-RFN) model material.

age accumulation only, effectively limiting crack growth and enhancing the resilience of the system at hand.

References

1. Gautieri A., Vesentini, S., Redaelli, A. and Buehler, M. J.: Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. Nano Lett. 11 757-766 (2011).
2. Gao, H.: Application of fracture mechanics concepts to hierarchical biomechanics of bone and bone-like materials. Int. J. Fracture 138, 101-137 (2006).
3. Lennartz-Sassinek, S., Zaiser, M., Main, I. G., Manzato, C. and Zapperi, S.: Emergent patterns of localized damage as a precursor to catastrophic failure in a random fuse network. Phys. Rev. E 87, 042811 (2013).
4. Zaiser, M., Lennartz-Sassinek, S. and Moretti, P.: Crack phantoms: localized damage correlations and failure in network models of disordered materials. J. Stat. Mech. P08029 (2015).
5. Gao, H.; Wang, X.; Yao, H.; Gorb, S.; Arzt, E.: Mechanics of hierarchical adhesion structures of geckos. Mechanics of Materials 37, 275285 (2005)
6. Yao, H., Gao, H.: Mechanics of robust and releasable adhesionin biology: Bottomup designed hierarchicalstructures of gecko. J. Mech. Phys. Solids 54 , 11201146 (2006)
7. Moretti, P., Dietemann, B., Esfandiary, N., Zaiser, M.: Avalanche precursors of failure in hierarchical fuse networks. Sci. Rep. 8, 12090 (2018)
8. de Arcangelis, L., Redner, S. and Herrmann, H. J.: A random fuse model for breaking processes. J. Phys. Lett. 46, L585 (1985).
9. Alava, M. J., Nukala, P. K. V. V. and Zapperi, S.: Statistical models of fracture. Adv. Phys. 55, 349-476 (2006).
10. Barraclough, T.W., Blackford, J.R., Liebenstein, S., Sandfeld, S., Stratford, T.S., Weinlnder, G. and Zaiser, M.: Propagating compaction bands in confined compression of snow Nature Phys. 13, 272275 (2017).
11. Zapperi, S., Nukala, P. K. V. V. and Simunovi S.: Crack roughness and avalanche precursors in the random fuse model. Phys. Rev. E 71, 026106 (2005).

Persistence of hierarchical network organization in biological systems

Ali Safari[1], Miguel Ángel Muñoz[2], and Paolo Moretti[1]

[1] Institute for Materials Simulation,Fredrich-Alexander-University Erlangen-Nuremberg, Dr-Mack-Str 77 90762 Fürth, Germany,
ali.s.safari@fau.de,
http://www.matsim.techfak.uni-erlangen.de/staff/ali-safari.shtml
[2] Departamento de Electromagnetismo y Fsica de la Materia e Instituto Carlos I de Fsica Terica y Computacional, Universidad de Granada, Granada E-18071, Spain

1 Introduction

Human brain networks are well known examples of biological systems, which exhibit a hierarchical modular structure [1, 2]. Collagen based biological matter, including bone and tendon, share the same hierarchical organization. The ubiquity of hierarchical organization in biological systems is often ascribed to the enhanced resilience that the hierarchical organization brings about.

In this work we address the problem of persistence of the hierarchical organization upon variation of relevant system parameters. Can we highlight universal measures of the hierarchical organization of a biological system? How robust are these measures against parameter changes? While there is a general agreement about the hierarchical nature of brain organization, their properties seem to depend on the correlation thresholds applied to the dense correlation matrices as well as on the nature of the functional process hosted by the network (e.g. subcritical vs. supercritical). Similarly, hierarchically organized biomaterials undergo significant changes under load, as damage (the number of broken links) advances [3]. Is a damaged hierarchical material still hierarchical?

2 Results

In order to verify to which extent small spectral gaps extend to functional connectivity, we generate functional netwoks by computing coactivation matrices, as suggested in [4], simulation spreading dynamics on hierarchical modular networks (HMN) [2, 5]. Spreading dynamics is associated with a spreading rate λ, which can be below or above a critical λ_c. We generating dense coactivation matrices in both cases, we apply varying finite and positive thresholds T and extract the sparse adjacency (or weight) matrix of the functional network. Upon increasing T in Figure 1 , in the subcritical case, the spectral gap drops by more than two orders of magnitude well before the network fragments. The functional network indeed inherits the small-gap property of the structural one that generated it. In other words, it is hierarchical too. In the supercritical case, instead, the two transitions become closer and, more importantly, as soon as the spectral gap starts

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

decreasing, the giant connected components shrinks too. In this case the functional network never exhibits a small spectral gap except when it is fragmented: the functional network is not hierarchical. Figure 1 also shows how this clear-cut separation between subcritical and supercritical is reflected by the degree distributions of the ensuing functional networks, which exhibit exponential tails in the subcritical case (mimicking the degree distribution of the underlying structural network), and heavy power-law tails in the supercritical regime. We show that in the case of functional brain connectivity, the hierarchical organization is persistent upon variation of thresholds in the subcritical and near-critical state that is commonly associated with healthy brain function, and is lost in the supercritical regime which is often associated with pathological conditions such as epilepsy. We observe the same persistent behavior in the context of fracture of

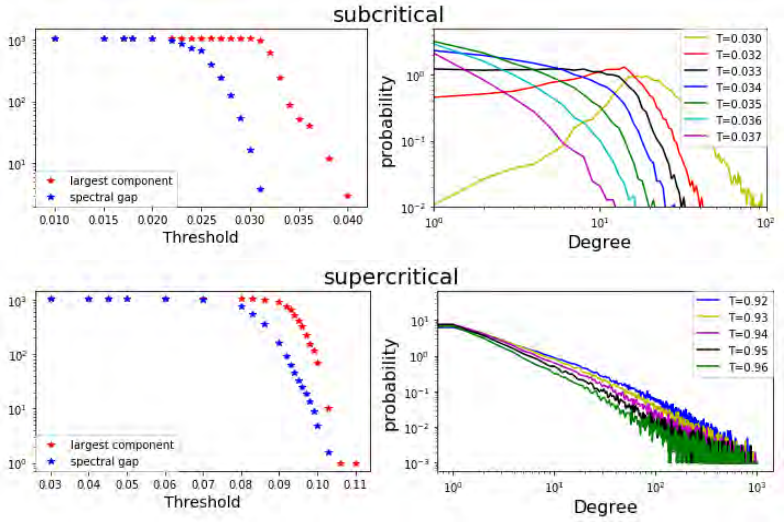


Fig. 1. comparisons between spectral gaps and giant connected component sizes (left) and Degree distribution (right) for functional coactivation networks generated from HMNs of size $N = 1024$.

hierarchically patterned network models of biological materials using the Hierarchical Fuse Network model (HFN) [3]. The critical point for such systems is represented by the peak load, the maximum amount of mechanical stress such systems can withstand. We measure eigenvector localization as an indicator of the hierarchical organization of such networks as damage progresses, quantified as the inverse participation ratio (IPR) of the eigenvectors corresponding to eigenvalues in the lower spectral edge of the network laplacian. We also consider the case of a non-hierarchical reference square lattice for comparison. As it is shown in Figure 2 top, in hierarchical systems, IPR values always increase with damage and the corresponding eigenvalues always decrease. In Figure 2 bottom, the IPR determined for a typical eigenvector increases exactly at the peak current, providing a clear-cut indicator of system's incipent failure. Given the rescaling on the vertical axis of Figure 2 bottom, we can conclude that damage induced

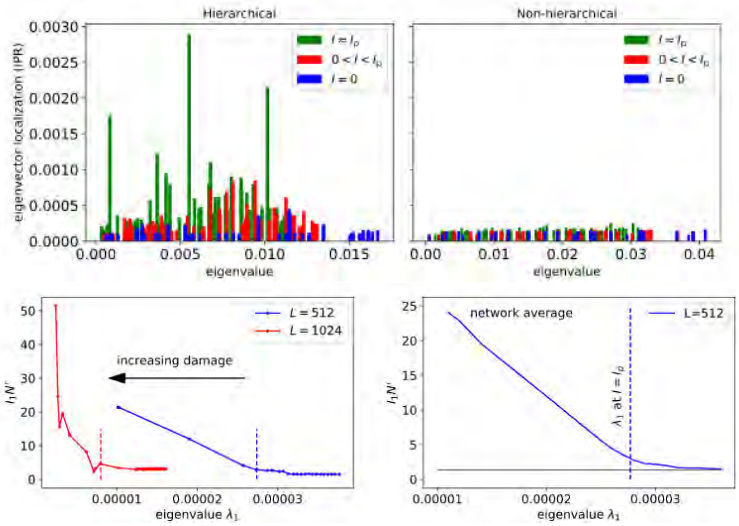


Fig. 2. Eigenvector localization in a hierarchical fuse network (top left) and in a reference square lattice (top right), at three different load stages and Evolution of localization in large-scale hierarchical fuse network under load (bottom).

localization is a robust and persistent phenomenon in hierarchical systems, while in non-hierarchical systems it only appears at the peak load. The robustness of our results is further confirmed in Figure 2 (bottom right), where the above results are averaged over different network realizations.

Summary. We study the persistence of hierarchical organization against structural changes, in network models of biological relevance. We focus on two examples, functional brain networks and network models of collagen based biological materials. We quantify the hierarchical organization by looking at specific spectral properties such as spectral gaps and eigenvector localization. We find that in both cases, normal function is associated with persistent hierarchical traits that do not depend on parameter variation or damage accumulation.

References

1. Sporns, O., Tononi, G., Kötter, R.: The Human Connectome: A Structural Description of the Human Brain. PLoS Comput. Biol. 1(4), e42 (2005)
2. Moretti, P., Muoz, M. A.: Griffiths phases and the stretching of criticality in brain networks. Nature communications. 4, 2521 (2013)
3. Moretti, P., Renner, J., Safari, A., Zaiser, M.: Graph theoretical approaches for the characterization of damage in hierarchical materials. EPJB. 92(5), 97 (2019)
4. Hütt, M. T., Kaiser, M., Hilgetag, C. C. Perspective: network-guided pattern formation of neural dynamics. Phil. Trans. R. Soc. 369(1653), 20130522 (2014)
5. Safari, A., Moretti, P., Muñoz, M. A. Topological dimension tunes activity patterns in hierarchical modular networks. New Journal of Physics. 19(11), 113011 (2017)

Reachability Analysis in Discrete State Reaction Networks with Conservation Laws

Gergely Szlobodnyik[1] and Gábor Szederkényi[1,2]

[1] Faculty of Information Technology and Bionics, Pazmany Peter Catholic University, Prater u. 50/a, H-1083 Budapest, Hungary,

[2] Systems and Control Laboratory, Institute for Computer Science and Control (MTA SZTAKI) of the Hungarian Academy of Sciences, Kende u. 13-17, H-1111 Budapest, Hungary

1 Introduction

Dynamical behavior of complex systems of several interacting components can be modeled by either continuous or discrete state models [1]. If the amount of the interacting components is high, it is reasonable to use continuous differential equation based modeling approaches to characterize the dynamical behavior of the system. However, in the case of low abundance of the interacting components, it is worth introducing discrete state models capable of quantitatively describing the discrete state evolution of all the components. In the latter discrete state case, the so-called reachability problem is strongly related to the quantitative dynamical behavior encoded by the underlying network topology: given an initial state X_0 and a target state X', is it possible to reach X' from X_0 along a finite non-negative state transition sequence? Through reachability analysis several problems of great importance can be analyzed, such as the existence of malicious states in a biochemical system or extinction events, i.e. state transition sequences resulted in the lack of cretain components.

In this paper we consider discrete state Chemical Reaction Networks (d-CRNs), a commonly used modeling approach employed for (bio)chemical systems when the molecular count of the species is low (e.g. < 100 molecules). We discuss subclasses of d-CRNs obeying conservation laws. We provide a set of conditions under which the reachability realiation is equivalent to the existence of a non-negative integer solution of the respective d-CRN state equation. We show how the results can be used in practice to efficiently analyse the dynamical behavior of d-CRNs in terms of the decidability of reachability problems. Our findings are shown on a representative example.

2 Results

A discrete state Chemical Reaction Network (d-CRN) can be described by a triple $(\mathscr{S},\mathscr{C},\mathscr{R})$ so that:

$$\mathscr{S} = \{s_i \mid i \in \{1, \dots, n\}\}$$
$$\mathscr{C} = \{y_j = \sum_{i=1}^{n} \alpha_{ji} s_i \mid \alpha_{ji} \in \mathbb{Z}_{\geq 0},\ j \in \{1, \dots, m\},\ i \in \{1, \dots, n\}\}$$
$$\mathscr{R} = \{(y_i, y_j) \subset \mathscr{C} \times \mathscr{C} \mid i \neq j\}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where s_i is the ith **species** and y_j is the jth **complex** of the network. α_{ij} is the **stoichiometric coefficient** of the ith species in the jth complex. A reaction $y_i \to y_j$ with **source complex** y_i and **product complex** y_j is represented by an ordered pair (y_i, y_j).

A directed graph $G = G(V,E)$ can be uniquely associated to a d-CRN so that the edge set E and vertex set V correspond to the complex set $\mathscr{C}$ and the reaction set $\mathscr{R}$, respectively. Two nodes $v_i \in V$ and $v_j \in V$ are connected by a directed edge $e \in E$ pointing from v_i to v_j iff $y_i \in \mathscr{C}$ and $y_j \in \mathscr{C}$ have a common reaction $(y_i, y_j) \in \mathscr{R}$. From now on under the term structure we mean the topology of the directed graph representation of the examined d-CRN.

The **stoichiometric matrix** $\Gamma \in \mathbb{Z}^{n\times m}$ of $\mathscr{N}$ is defined as

$$\Gamma = [r_1 \ \dots \ r_m] \tag{1}$$

where r_i is the ith **reaction vector**, i.e. $[r_i]_j$ gives the number of molecules of species $s_j \in \mathscr{S}$ produced or consumed by the ith reaction. For each reaction vector r_i we can introduce the vectors $\overline{y_i^-}$ and $\overline{y_i^+}$ so that $[\overline{y_i^-}]_j$ and $[\overline{y_i^+}]_j$ denote the stoichiometric coefficient of the jth species in the ith reaction's source complex and product complex, respectively. We also introduce the matrix Γ^- as follows:

$$\Gamma^- = [\overline{y_1^-} \ \dots \ \overline{y_m^-}]^\top.$$

We note that a pair (Γ, Γ^-) uniquely characterizes the underlying reaction network structure.

In this paper we restrict our attention to the subclasses of sub-and superconservative reaction networks. A d-CRN of stoichiometric matrix $\Gamma \in \mathbb{Z}^{n\times m}$ is said to be **subconservative** (**superconservative**), if there exists a strictly positive real-valued vector z of dimension m, so that $z^\top \Gamma \leq 0$ ($z^\top \Gamma \geq 0$).

A state $X^{'} \in \mathbb{Z}^n_{\geq 0}$ is said to be **reachable** from a state $X \in \mathbb{Z}^n_{\geq 0}$ (denoted by $X \rightsquigarrow X^{'}$) if there exists a path in the state space so that $X = X_{\nu(1)} \to X_{\nu(2)} \to ... \to X_{\nu(l)} = X^{'}$. The associated state tranisition sequence is denoted by $\sigma_X = X_0 \dots X^{'}$.

Problem statement: consider a d-CRN $\mathscr{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$ of stoichiometric matrix $\Gamma_{\mathscr{N}}$ and two integer states $X_0,\ X^{'} \in \mathbb{Z}^n_{\geq 0}$. Is it possible to find a non-negative state transition sequence $\sigma_X = X_0 \dots X^{'}$ [4]?

A necessary condition of the above reachability problem is the existence of a non-negative integer $c \in \mathbb{Z}^m_{\geq 0}$ solution of the respective d-CRN state equation:

$$\Gamma_{\mathscr{N}} c = X^{'} - X_0, \qquad c \in \mathbb{Z}^m_{\geq 0}. \tag{2}$$

E.q. (2) implies an integer programming problem which in general requires the introduction of additional supplementary variables [2]. This may lead to computational intractability. This problem motivates us to seek network topology related conditions under which E.q. (2) is a sufficient and necessary condition of the reachability relation.

Before we state our main result the follwoing supplementary variable is introduced:

$$[M(\Gamma^-)]_i = max\Big\{[\Gamma^-]_{ij} : j = 1,\dots,m\Big\}, \qquad i = 1,\dots,n. \tag{3}$$

Proposition 1 *[3] Let us consider d-CRN $\mathscr{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$ with stoichiometric matrix $\Gamma \in \{-1,0,1\}^{n\times l}$ and $\Gamma^{-} \in \{0,1\}^{n\times l}$ for which $\mathscr{C} = \mathscr{S}$. Assume that $\mathscr{N}$ is either subconservative or superconservative. Assume that for all $r \in \mathscr{R}$ reactions $\sum_{i=1}^{n} [\overline{y}^{+}]_i = 1$ and $\sum_{i=1}^{n} [\overline{y}^{-}]_i \leq 1$ hold. Let us consider two states X_0, $X^{'} \in \mathbb{Z}^{n}_{\geq 0}$ so that $X_0 \succeq M$ and $X^{'} \succeq M$ hold where $M = M(\Gamma^{-})$. Then*

$$X_0 \rightsquigarrow_{\mathscr{N}} X^{'} \iff \exists c \in \mathbb{Z}^{l}_{\geq 0} : \; X_0 + \Gamma c = X^{'} \tag{4}$$

For more details on the above theoretical findings the reader is refereed to [3].

Figure 1 depicts a representative example of a reaction network structure satisfying the conditions of Proposition 2.

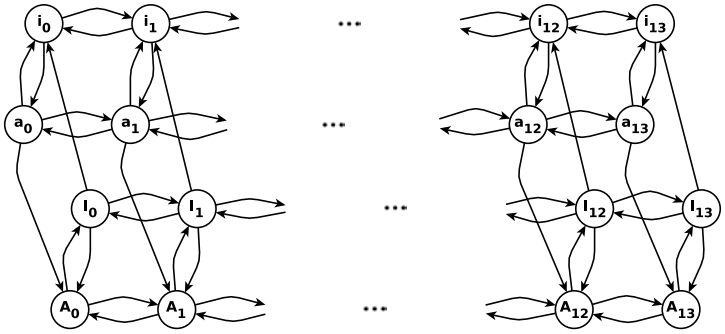


Fig. 1. Nuclear factors of activated T-cells (NFAT) are transcription factors that can exist in both highly phosphorylated and dephosphorylated states [5]. The transition between active and inactive states of the protein is regulated by the level of phosphorylation. Lower case letters denote the protein located in the cytoplasm while upper case letters refer to the protein in the nucleus. a_j, A_j and i_j, I_j for $j = 0, \ldots 13$ denote the active and inactive proteins, respectively. Lower indices denote the number of phosphorylated residues.

References

1. T. G. Kurtz, “The relationship between stochastic and deterministic models for chemical reactions”, The Journal of Chemical Physics, vol. 57, no. 7, pp. 2976–2978, 1972.
2. G. Szlobodnyik, G. Szederkényi, M. Johnston, ”Reachability analysis of subconservative discrete chemical reaction networks”, MATCH Commun. Math. Comput. Chem., 81(3), pp. 705-736., 2019.
3. G. Szlobodnyik, G. Szederkenyi, ”Reachability Analysis of Low-Order Discrete State Reaction Networks Obeying Conservation Laws”, Complexity, Vol. 2019, Article ID 1035974, 13 pages, 2019.
4. S. R. Kosaraju, “Decidability of reachability in vector addition systems”, in STOC, pp. 267–28, ACM, 1982.
5. C. Salazar and T. Höfer, “Allosteric regulation of the transcription factor NFAT1 by multiple phosphorylation sites: a mathematical analysis,” Journal of Molecular Biology, vol. 327, no. 1, pp. 31–45, 2003.

Relation between connectivity and coupling in the Chilean subduction zone: a first approach

Fernanda Martin[1] and Denisse Pastén[1]

Universidad de Chile, Chile,
denisse.pasten.g@gmail.com

1 Introduction

We have applied the theory of complex networks to earthquakes, characterizing the complete Chilean subduction zone with parameters of complex networks, for this purpose, we have built a time-based complex network [1–3], making cells of side size δ in the zone under study. If one of these cells contains an earthquake, we call this cell as a node. The network is built following the time occurrence of the seismic events, i.e., the nodes connect with each other follow the temporal sequence of seismic events the nodes [3, 4].

For this study we have divided the subduction zone of Chile into regions, from the northern zone to the southern zone, each region is 300 km long. The time-based complex network was built in each region. The seismic data set analyzed were collected between January 2005 and March 2017 by the National Seismological Centre of Chile (Servicio Sismológico Nacional, CSN [7]), so we have a completeness data set with a total number of 38 083 seismic events measured along the Chilean coast, from 17.9° to 39.1° South Latitude and between 67.5° and 75° West Longitude. The magnitude of completeness is M_w3.0 for all the data set. The data set used in this analysis could be found and downloaded in www.sismologia.cl [7].

2 Results

We compute the critical exponent g from the probability distribution of connectivity ($P(k) \sim k^{-\gamma}$) and the Average Shortest Path Length (ASPL) for 22 regions along Chilean coast. We compare these results against the average coupling of the tectonics plates, because the subduction is the physical mechanism that induces the earthquake occurrence in Chile and the coupling and stress play an important role in the in this occurrence, in order to looking for some connection between the physical parameters related to the occurrence of seismic events and complex networks. The results are shown in Figs. 1, 2 and 3.

Fig. 1 shows the spatial evolution of the parameter γ between 2005 and 2017 along the Chilean coast. We can observe a change of this parameter, in the northern and the southern zone the value of γ is lower than the central zone. Fig. 2 shows the spatial evolution of the ASPL from the northern zone of Chile to the southern zone of Chile, as Fig. 1, this value changes in each window studied. In Fig. 3 we can observe the agreement between the coupling between Nazca plate and South American plate and the value of the critical exponent γ.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

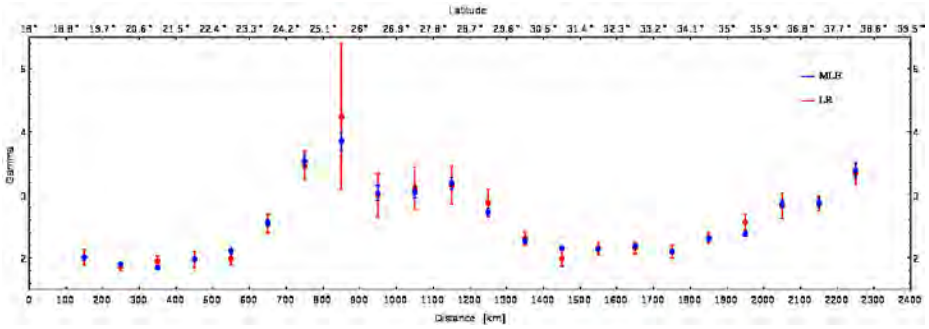


Fig. 1. Value of the critical exponent γ for the probability distribution of connectivity, computed along the Chilean coast, from the 18° to the 39° South Latitude.

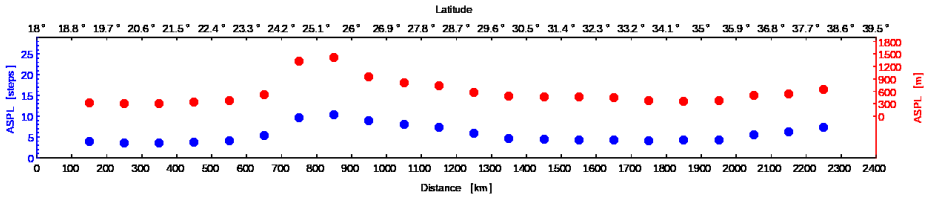


Fig. 2. Value of the ASPL for each region of 300 km, between the northern zone of Chile and the southern.

3 Conclusions

In this work we present a first effort to find a relationship between the parameters of complex networks, as the critical exponent γ of the probability distribution of connectivity ($P(k) \sim k^{-\gamma}$) and the physical dynamics involved in the earthquake occurrence. Fig. 1 shows how the critical exponent γ change along the Chilean coast. This exponent has its greatest value in the central north zone of Chile, and decreases in the northern and the southern zone. The Average Shortest Path Length (ASPL) has a similar behavior than the exponent γ, the greatest value is in the central north zone of Chile, Fig. 2.

Fig. 3 shows the average coupling measured by Métois et al. [5, 6] versus the values of the critical exponent γ. Fig. 3 suggests an agreement between these two parameters in the central zone of Chile, but it is possible to observe a disagreement between the values of these two parameters in the northern zone and the southern zone of Chile. Another important fact to consider is the occurrence of three large earthquakes in Chile during the time analyzed. The M_w 8.8 Maule megathrust in 2010 (southern Chile), with a rupture zone of 450 km, the M_w8.2 Iquique earthquake in 2014 (northern Chile), with a rupture zone of 150 km and the M_w8.3 Illapel earthquake in 2015 (central-north Chile) with a rupture zone of the 200 km. If we consider the effect of these three mega earthquakes, we could suggest a relation between γ and the occurrence of a large earthquake. The epicenter of Maule megathrust it was at 36.2°, in Fig. 3 we can observe a growth of

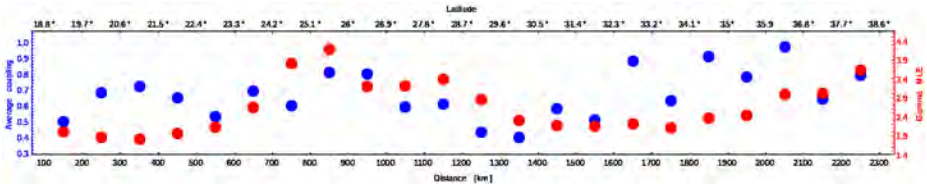


Fig. 3. Average coupling in the subduction zone between Nazca plate and South American plate, versus the critical exponent g along Chilean coast.

γ in the southern area of the epicenter. For the large earthquake in Iquique, the epicenter it was located at 19.5°, with a similar trend for the value of γ. Finally, the epicenter of Illapel large earthquake is located at 31.5°.

The main goal of this analysis is the proposal of a relation between the occurrence of a large earthquake and a change in the value of the critical exponent γ.

This is a first approach to try to connect the critical exponent g with the physical dynamics of the subduction mechanism of earthquake occurrence.

References

1. Abe, S., Suzuki N.: Complex-Network Description of Seismicity. Nonlinear Proc. Geophys. 13, 145–150 (2006)
2. Abe, S., Pastén, D., Muñoz, V., Suzuki, N.: Universalities of Earthquake-Network Characteristics. Chinese Science Bulletin 56, 3697–3701 (2011)
3. Abe, S., Pastén, D., Suzuki, N.: Finite data-size scaling of clustering in eartquake networks. Physica A 390, 7–11 (2011)
4. Pastén, D., Torres, F., Toledo, B., Muñoz, V., Rogan, J., Valdivia, J.A.: Time-Based Network Analysis Before and After the M_w 8.3 Illapel Earthquake 2015 Chile 173, 2267–2275 (2016)
5. Métois, M., Valderas-Bermejo, M.C., Ortega, I., Socquet, A., Vigny, C., Carrizo, D., Peyrat, S., Delorme, A., Maureira, E.: Revisiting the North Chile seismic gap segmentation using GPS-derived interseismic coupling. Geophys. J. Int. 194, 1283–1294 (2013)
6. Métois, M., Vigny, C., Socquet, A.: Interseismic coupling, megathrust earthquakes and seismic swarms along the Chilean subduction zone 38°-18°S. Pure and Applied Geophysics 173, 1431–1449 (2016).
7. Web site of Servicio Sismológico Nacional (Chile), http://www.sismologia.cl/

Algorithmically identifying dynamical subsystems of genetic-metabolic networks in bacteria

Santhust Kumar[1], Saurabh Mahajan[2], and Sanjay Jain[3,4]

[1] Jacobs University, Bremen 28759, Germany,
s.santhust@jacobs-university.de,
[2] St. Joseph's College (Autonomous), Bangalore, India 560027,
[3] Department of Physics and Astrophysics, University of Delhi, Delhi 110007, India,
[4] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

1 Introduction

The process of genetic regulation is often influenced by other processes in a cell, for e.g., metabolism. The structure of genetic regulatory networks (GRN) at a system level, however, has often been studied in isolation. We study the architecture of GRNs of two bacteria, *Escherichia coli* and *Bacillus subtilis*, under the effect of feedback from respective metabolic networks (MN). We organize the GRNs in a causal flow-of-command hierarchy and show that the inclusion of feedbacks from MN into GRN greatly alters this causal flow of information in the GRNs. Through a combination of graph theoretic approach of finding modules *via* strongly connected components (SCCs) [1] and computational functional approach of flux balance analysis (FBA) for simulating growth through metabolic models, we further show that the SCCs of the GRN augmented with feedbacks from MN can be considered as modules or subsystems with logically relatable and biologically relevant functions.

2 Results

2.1 Feedback into GRN from MN

The GRN and MN of *E. coli* and *B. subtilis* were obtained from publicly available databases and previous published works [2–5]. We identify the metabolites in metabolic network which can form complexes with transcription factors (TFs) in GRN, and then use this information to elucidate feedbacks from metabolic network into different levels of a hierarchically organized GRN in which all regulations point downward [6], Fig. 1A.

2.2 Hierarchical structure of GRN augmented with feedback from metabolic network

We obtain functional feedbacks from metabolic network into the GRN by choosing metabolites from reactions deemed essential by flux balance analysis upon simulation of growth in minimal media environmental conditions (ECs)—*E. coli*: 158 ECs (89 aerobic, 69 anaerobic), *B. subtilis*: 212 ECs (all aerobic). Using these feedbacks from

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

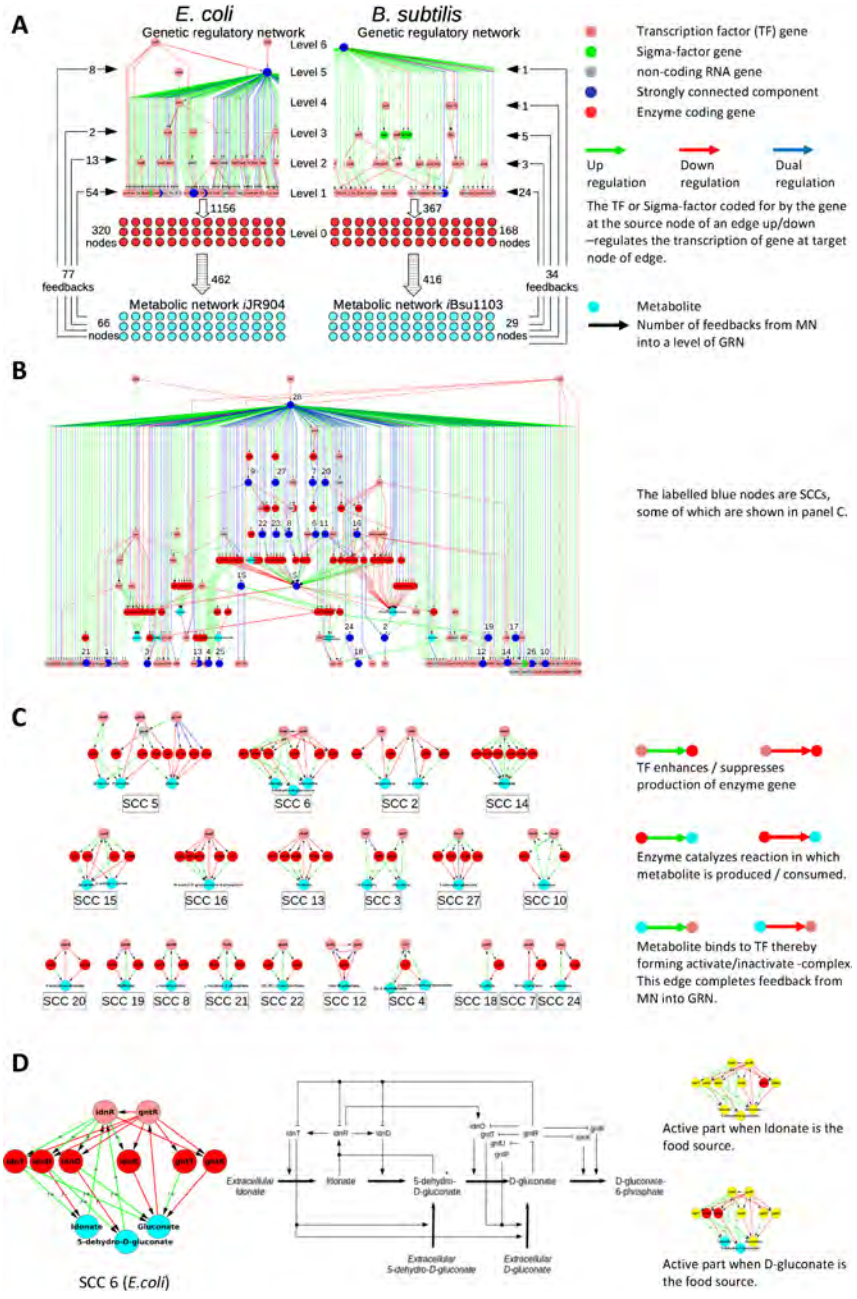


Fig. 1. Feedbacks, hierarchical structure and modules in GRNs with feedback from metabolic network. (A) Schematic of feedback from metabolic network into different levels of GRN. **(B)** Hierarchical structure of the GRN of *E. coli* augmented with functional feedbacks from metabolic network (graph $\mathscr{G}$). The blue nodes represent the SCCs/modules, shown in next panel. **(C)** Strongly connected components (SCCs) from GRN augmented with feedback from metabolic network for *E. coli*. Their location in the hierarchy is shown in panel B. (A few SCCs of small size have been omitted from the figure due to space constrains, for details and for *B. subtilis* see [6]). **(D)** An example of a module from *E. coli* along with its proximal circuit diagram and its activity under two simulated environments. In the proximal circuit diagram a pointed arrow implies activation/production/up-regulation and a blunt arrow implies deactivation/consumption/down-regulation, depending upon whether the regulated node is a TF, metabolite or enzyme, respectively. Nodes highlighted in yellow indicate active part of the module in respective environmental conditions.

metabolic netowork, we augment the GRN and obtain graph $\mathscr{G}$. Next, we organize the augmented GRN into a causal hierarchical organization by applying graph condensation and iterative leaf removal algorithms on graph $\mathscr{G}$ [6]. The hierarchical structure of *E. coli* graph $\mathscr{G}$ is shown in Fig. 1B. The hierarchical structure of $\mathscr{G}$ is very different from that of GRN without inclusion of feedback in that the causal flow of information is significantly altered. The blue nodes represent strongly connected components (SCCs): 28 SCCs for *E. coli* and 14 SCCs for *B. subtilis*. Many strongly connected components of *E. coli* for graph $\mathscr{G}$ are shown in Fig. 1C, (see [6] for full list). The blue node at the top of the hierarchical structure is the largest SCC (LSCC), which has a size of 97 nodes for *E. coli* (13 for *B. subtilis*). The largest strongly connected component has a more complex structure and requires further study.

2.3 Regulatory modules in GRN augmented with feedbacks from metabolic network

Next, we study the activity of all strongly connected components (SCCs) of graph $\mathscr{G}$ in each of the simulated environmental conditions (ECs) by carefully developing proximal 'circuit diagrams' around the nodes of these SCCs. We find that most of the strongly connected components can be ascribed biologically relevant functional roles and the proximal circuit diagram relates well to their activity in different environmental conditions under elementary on/off -logic. An example of this is given in Fig. 1D through the Idonate-Gluconate module in *E. coli* which is partially active in right manner for the uptake of Idonate or Gluconate as food in their respective simulated minimal media environmental condition. For a complete list of modules, ascribed biological function and corresponding proximal elementary circuit diagrams, for both *E. coli* and *B. subtilis*, see [6].

Summary. We studied the architecture of gene regulatory network (GRN) under the effect of feedback from metabolic network (MN) and present an updated hierarchical structure thereby showing that the inclusion of feedbacks from metabolic network into the GRN significantly alters the causal flow of information in the GRNs. We algorithmically identify dynamical sub-systems of the joint genetic-metabolic network and show that the identified modules posses biologically relevant and logically relatable functionality. The list of identified modules obtained in our work may be used in future as a starting point for improved modeling of sub-systems in these bacteria. Further, our algorithmic approach may be automated to find important sub-systems in other organisms as and when their GRN and MN become available.

References

1. Rodrguez-Caso, C., Corominas-Murtra, B., Sol, R.V.: On the basic computational structure of gene regulatory networks. Molecular BioSystems **5**(12) (November 2009) 1617–1629
2. Salgado, H., Peralta-Gil, M., Gama-Castro, S. *et al*: RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Research **41** (2013) D203–D213

3. Reed, J., Vo, T., Schilling, C., Palsson, B.: An expanded genome-scale model of *Escherichia coli* k-12 (*i*JR904 GSM/GPR). Genome Biology **4**(9) (2003) R54
4. Freyre-Gonzalez, J., Manjarrez-Casas, A., Merino *et al*: Lessons from the modular organization of the transcriptional regulatory network of *Bacillus subtilis*. BMC Systems Biology **7**(1) (2013) 127
5. Henry, C.S., Zinner, J.F., Cohoon, M.P., Stevens, R.L.: *i*Bsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. Genome Biology **10**(6) (2009) R69
6. Kumar, S., Mahajan, S., Jain, S.: Feedbacks from the metabolic network to the genetic network reveal regulatory modules in *E. coli* and *B. subtilis*. PLOS ONE **13**(10) (October 2018) e0203311

Gene coexpression networks for the study of *Rhizobium leguminosarum*

Javier Pardo-Diaz[1,2], Mariano Begueressi-Diaz[3], Phillip Poole[2], Charlotte M Deane[1], and Gesine Reinert[1]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK,
jdiaz@stats.ox.ac.uk,
WWW home page: https://www.stats.ox.ac.uk/~ jdiaz/
[2] Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK,
[3] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

1 Introduction

In this work, we aim to generate gene coexpression networks in which nodes correspond to the genes and edges represent high positive correlations in their expression across different samples (6). Therefore, genes which are expressed under the same conditions are connected in the network, easing the study and visualisation of the expression data (8, 11). Genes that are coexpressed across multiple conditions are likely to have related functions (4, 9, 10). This makes possible to deduce gene function using *guilt by association* approaches. Network-based methods such as community detection may help in this process. This procedure is especially useful if the studied organism is poorly annotated.

The most commonly used methods to generate gene coexpression networks are based on the absolute value of the Pearson correlation coefficient of the expression of each pair of genes (11). Based on this value, there are two possibilities: treating the coexpression as a continuous value and constructing weighted networks; or constructing an unweighted network by applying a threshold. The later approach allows the selection only the strongest relationships but may give rise to a loss of information.

We use *Rhizobium leguminosarum* gene expression data from a collection of microarrays to generate both unweighted and weighted coexpression networks. *R. leguminosarum* is an α-proteobacterium that fixes atmospheric nitrogen when associated with legumes (eg. peas, beans, lentils). *R. leguminosarum* transforms molecular nitrogen into ammonia which can be assimilated by plants. Nitrogen fixation improves the growth of plants as nitrogen is one of the limiting factors during the growth process (3). *R. leguminosarum* experiences very large changes in its metabolism from the *free-living* bacteria to the *plant-associated* bacteria (12). These changes are reflected in the gene expression levels (5) and we aim to detect them in our gene coexpression networks.

2 Results

We present a pipeline to generate and study gene coexpression networks from gene expression data (Fig. 1). Firstly, we set the expression of the 20% lowest expressed genes

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

from each microarray to zero to remove noise that may compromise later steps in our network construction methodology. We also apply the quantile normalisation method (2) to our expression matrix to make the distribution of the expression values of each experiment identical in statistical properties. This allows us to compare values from different microarrays. After the preprocessing steps, we calculate the correlation between the expression of each pair of genes in the expression matrix to obtain a symmetric correlation matrix.

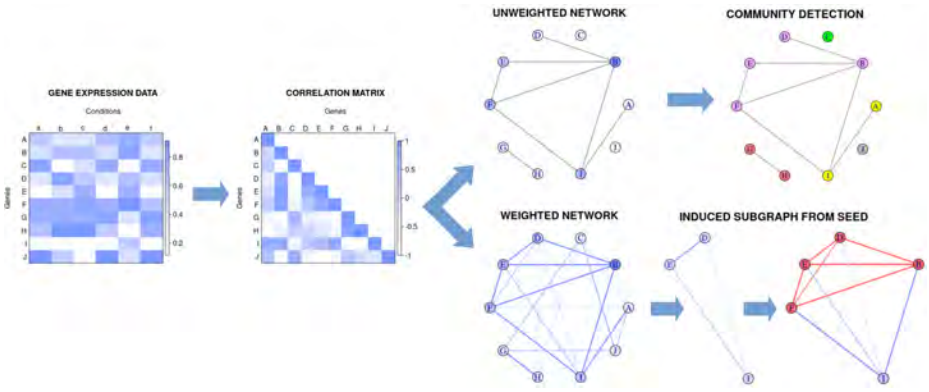


Fig. 1: Pipeline to generate gene coexpression networks. The correlation matrix is obtained from the gene expression data. Afterwards, there are two possibilities: generating an unweighted network and use community detection to find functionally-related genes, or generating a weighted network and use an ego-network based approach.

We impose a threshold to the correlation matrix to obtain an unweighted network network with edges only between the genes whose expression correlation is higher than the threshold. We test different thresholds and score the resulting networks using a Monte Carlo test and metabolic information from biological databases. The optimal threshold (0.63) balances noise reduction whilst retaining functional information. In this network, with density 1.2%, 82% of the top 500 pairs genes reported to be coexpressed according to the database STRING are connected. To optimise the network partitions, we use the Louvain method (1), using 101 different resolution parameter values. We study communities enriched in genes which are involved in the same biological process, restricting our evaluation to only those with between 6 and 60 nodes since sizes outside this range are not interesting from a practical point of view (7). We find that genes involved in the same metabolic pathways tend to be in the same communities. It would be interesting to study other methods of association and different community detection algorithms to assess performance and robustness of the network.

Alternatively, the correlation matrix can be used to generate a weighted network. In this case, the weights of the edges between pairs of genes are the values of the correlation of their expression. We use only positive correlation values greater than 0.3. We use an ego-network based approach to obtain the genes related to a given set of

genes. This approach allows us to recover half of the genes annotated as ribosomal genes in *R. leguminosarum* by using the other half of those genes as seed.

The next step in our analysis will be the use of RNAseq data from transcriptional regulators mutants. This information will allow us to study the role of those proteins in the coexpression network.

Summary We have applied our pipeline to generate unweighted and weighted *R. leguminosarum* gene coexpression networks. Our results suggest that both such networks can be a useful guide in the identification of genes involved in the same biological processes, in the prediction of gene function, and in the verification of genome annotations.

References

[1] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
[2] Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2), 185–193 (2003)
[3] Gutiérrez, R.A.: Systems biology for enhanced plant nitrogen nutrition. Science 336(6089), 1673–1675 (2012)
[4] Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al.: Functional discovery via a compendium of expression profiles. Cell 102(1), 109–126 (2000)
[5] Karunakaran, R., Ramachandran, V., Seaman, J., East, A., Mouhsine, B., Mauchline, T., Prell, J., Skeffington, A., Poole, P.: Transcriptomic analysis of Rhizobium leguminosarum biovar viciae in symbiosis with host plants Pisum sativum and Vicia cracca. Journal of Bacteriology 191(12), 4002–4014 (2009)
[6] Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., Pavlidis, P.: Coexpression analysis of human genes across many microarray data sets. Genome Research 14(6), 1085–1094 (2004)
[7] Luecken, M., Page, M., Crosby, A., Mason, S., Reinert, G., Deane, C.: CommWalker: correctly evaluating modules in molecular networks in light of annotation bias. Bioinformatics 34(6), 994–1000 (2017)
[8] Magwene, P.M., Kim, J.: Estimating genomic coexpression networks using first-order conditional independence. Genome Biology 5(12), R100 (2004)
[9] van Noort, V., Snel, B., Huynen, M.A.: Predicting gene function by conserved co-expression. TRENDS in Genetics 19(5), 238–242 (2003)
[10] Stuart, J.M., Segal, E., Koller, D., Kim, S.K.: A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643), 249–255 (2003)
[11] Weirauch, M.T.: Gene coexpression networks for the analysis of DNA microarray data. Applied Statistics for Network Miology: Methods in Systems Biology 1, 215–250 (2011)
[12] Wheatley, R.M., Ramachandran, V.K., Geddes, B.A., Perry, B.J., Yost, C.K., Poole, P.S.: Role of O2 in the growth of Rhizobium leguminosarum bv. viciae 3841 on glucose and succinate. Journal of Bacteriology 199(1), e00572–16 (2017)

Curvature-based analysis of Directed Hypernetworks

Wilmer Leal[1,2], Marzieh Eidi[2], and Jürgen Jost[2,3]

[1] Bioinformatics group, Leipzig University, 04103 Leipzig, Germany
[2] Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
[3] The Santa Fe Institute, 1399 Santa Fe, New Mexico, USA
wleal@mis.mpg.de, meidi@mis.mpg.de, jjost@mis.mpg.de

1 Introduction

Networks encoding symmetric binary relations between pairs of elements are mathematically represented by (undirected) graphs. Graph theory is a well developed mathematical subject, but empirical networks are typically less regular and also often much larger than the graphs that are mathematically best understood. Several quantities have therefore been introduced to characterize the large scale behavior or to identify the most important vertices in empirical networks. As the crucial structure of a graph is, however, given by the set of its edges rather than by its vertices, we should systematically define and evaluate quantities assigned to the edges rather than to the vertices.
Curvature is a notion originally introduced in the context of smooth Riemannian manifolds to measure local or global deviation of a manifold from being Euclidean. Ricci curvature specifically, as a local measure, provides relatively broad information about the structure of positively curved manifolds. Therefore, there have been several attempts to discretize curvature notions to other settings such as cell complexes [5], graphs [4] and undirected hypergraphs [7] for obtaining similar results. By this discretizations they have been able to transfer some of the analytical or topological properties of original smooth curvatures to these discrete spaces [6]. For the directed hypergraph case, these curvatures were introduced recently and very little is known about their descriptive power. In this paper, we first present the results of our discretizations of Forman-Ricci [1] and Ollivier-Ricci [2] curvature notions, then, we show that they are powerful tools for exploring local properties of directed hypergraph motifs. To conclude, we carry out a curvature-based analysis of the metabolic network of *E. coli*.

2 Results

Forman-Ricci Curvature. The structure of a graph is given by its edges. Therefore, a structural analysis of a graph should involve quantities describing local properties of edges, as a complement of the usual quantities of local properties of nodes. Forman-Ricci curvature serves that purpose. This notion was introduced by Forman for simplicial complexes and therefore, for graphs (graphs are one-dimensional simplicial complexes). Considering an undirected unweighted graph and an edge e with nodes i, j, it is simply given by $F(e) = 4 - \deg(i) - \deg(j)$. Edges connecting nodes with large degree have very negative Forman-Ricci curvature values, allowing a readily identification of those edges playing a key role in the cohesion of a network.
We generalize this notion to directed hypergraphs, [1]. Formally, a directed hypergraph is a couple $H = (V, E)$ where V is a set of vertices and E a set of ordered pairs of subsets

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

of V called hyperedges. Moreover, given a hyperedge $e = (e_i, e_j) \in E$, $e_i \subseteq V$ and $e_j \subseteq V$ are called the tail and head of e respectively. We define the Forman-Ricci curvature of e as

$$F(e) = |e_i| + |e_j| - \sum_{i \in e_i} \text{in-deg}(i) - \sum_{j \in e_j} \text{out-deg}(j) \tag{1}$$

Ollivier-Ricci Curvature. Similar to its smooth Riemannian counterpart, the definition of Ollivier-Ricci curvature involves comparing the average distance between the points of two balls (neighbourhoods) with the distance of their centers. In [2] we introduced this curvature notion for directed hypergraphs by using the Wasserstein distance between two probability measures associated to a directed hyperedge. We say that $u \to e_i$ if there exists a hyperedge $e = (e_k, e_i)$ such that $u \in e_k$. Similarly, $e_j \to v$ if there exists a hyperedge $e = (e_j, e_k)$ such that $v \in e_k$. Given a hyperedge $e = (e_i, e_j)$, we define two sets $\mathscr{M} := \{u : u \to e_i\} \cup \{i \in e_i : \text{there is no incoming hyperedge to } i\}$ called masses and $\mathscr{H} = \{v : e_j \to v\} \cup \{j \in e_j : \text{there is no outgoing hyperedge from } j\}$ called holes. Then, we assign a probability measure to each set, namely $\mu_{\mathscr{M}}$ and $\mu_{\mathscr{H}}$. For $u \in \mathscr{M}$ and $v \in \mathscr{H}$, we call $\mu_{\mathscr{M}}(u)$ and $\mu_{\mathscr{H}}(v)$ the size of the mass u and the size of the hole v, respectively. Considering the distance between each mass and each hole as the minimum number of directed hyperedges connecting them, this distance is at most 3. Now the question (formally called optimal transport problem) is how the first probability measure can be moved to the second one in an optimal way. We want to minimize expression (2) which iterates over all those matrices $\mathscr{E}$ (called transport plans) whose entries represent the amount of mass, out of $\mu_{\mathscr{M}}(u)$, to be moved from vertex u to vertex v, denoted by $\mathscr{E}(u,v)$. Moreover, $d(u,v)$ stands for their distance (with $d(u,v) \in \{0,1,2,3\}$).

$$\sum_{u \to e_i} \sum_{e_j \to v} d(u,v)\mathscr{E}(u,v) \tag{2}$$

Given an optimal transport plan, if m_x is the amount of mass that is moved at distance x, then the Ollivier-Ricci curvature κ of e is defined as $\kappa(e) = m_0 - m_2 - 2m_3$. It is bounded above by $\kappa = 1$ (reached when $m_0 = 1$ i.e. when each mass coincides with a hole of its same size) and below by $\kappa = -2$ (reached when $m_3 = 1$ i.e. each mass has to be moved at distance 3).

Connectivity motifs: Forman-Ricci vs Ollivier-Ricci. Fig. 1 shows the local structure of directed hypergraphs with positive, negative and zero values for both Ricci curvatures. For the given orange directed hyperedge e, $O(e)$ and $F(e)$ correspond to Ollivier and Forman curvatures respectively. Therefore, from left to right we can detect changes in the signs for Ollivier curvature while the sign of Forman is fixed. On the other hand, when we move vertically in the plot, Forman's sign change while Ollivier's sign is fixed. In the diagonal, directed hyperedges have the same sign for both curvatures.

Metabolism of *E.coli*. Fig. 2a) shows the number of metabolic reactions with $|e_i|$ reactants and $|e_j|$ products. 90% of chemical reactions have at most three reactants and three products (also observed for the whole Chemical Space [3]), which, according to equation 1, indicates that frequent curvature values in Fig. 2b) are ruled by the accumulated in- and out-degree. In particular, frequent values of curvature were found to distinguish bottle neck and redundant reactions in the metabolic network [1]. On the other hand, when considering the number of incoming neighbors of reactants and of

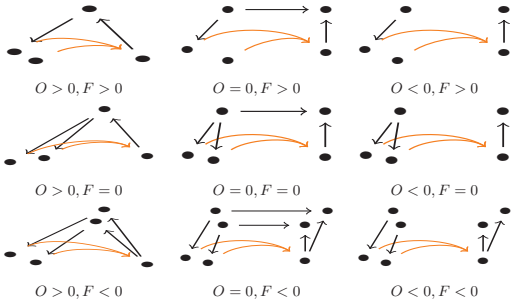


Fig. 1. Hypergraph motifs and their curvature sign. $F(e)$ is a balance between edge degree and node degree (in-degree for nodes in the tail and out-degree for nodes in the head), while $O(e)$ is a local measure of what is known in metabolic networks as topological overlap.

outgoing neighbors of products for every reaction, frequencies are of the order of hundreds and, for some reactions, almost the whole substrate set, as shown in Fig. 2c). The question that arises is how close are those masses and holes in the metabolic network. Ollivier-Ricci curvature distribution in plot Fig. 2d) gives us the answer: most masses and holes are at distance lower than 3, since the vast majority of them have curvature greater than -0.5. Less than 10% of incoming and outgoing neighbors are at distance 3. Only four reactions have curvature -2, indicating that their masses are at least three reactions away from their holes.

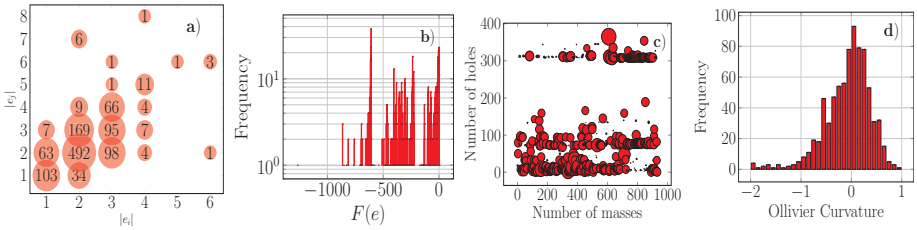


Fig. 2. a) number of reactants and products; b) $F(e)$ distribution; c) number of masses and holes; d) $O(e)$ distribution

References

1. Leal, W., Restrepo, G., Stadler, P. F., Jost, J.: Forman-Ricci curvature for Hypergraphs. arXiv e-prints, arXiv:1811.07825 (2018)
2. Eidi, M. & Jost, J.: Ollivier-Ricci curvature of Directed Hypergraphs. arXiv e-prints, arXiv:1907.04727 (2019)
3. Llanos, E. J., Leal, W., Luu, D.H., Jost, J., Stadler, P.F., Restrepo, G.: Exploration of the chemical space and its three historical regimes. Proceedings of the National Academy of Sciences of the United States of America 116 (26), 12660-12665 (2019)
4. Jost, J. & Liu, S.: Ollivier's Ricci curvature, local clustering and curvature-dimension inequalities on graphs. Discrete & Computational Geometry 51 (2), 300-322 (2014)
5. Forman, R.: Bochner's method for cell complexes and combinatorial Ricci curvature. Discrete & Computational Geometry 29 (3), 323-374 (2003)
6. Saucan, E.: Metric Curvatures and their Applications 2: Metric Ricci Curvature and Flow. arXiv e-prints, arXiv:1902.03438 (2019)
7. Asoodeh, S., Gao, T., Evans, J.: Curvature of Hypergraphs via Multi-Marginal Optimal Transport. In: 2018 IEEE Conference on Decision and Control (CDC). IEEE. 1180-1185 (2018)

Interacting gene networks poised at the edge of chaos?

Johan Dubbeldam[1]

Delft University of Technology, DIAM, Van Mourik Broekmanweg 6, Delft 2628 XE, The Netherlands,
j.l.a.dubeldam@tudelft.nl,
WWW home page: http://ta.twi.tudelft.nl/dv/users/dubbel

1 Introduction

A key, and often debated, concept in the physics of life is the criticality of living systems. The idea first put forward by Kaufman [1, 5] is that life evolves on the "edge of chaos", that is, a living system is sufficiently stable to sustain its organization, but has the ability to easily adapt itself to changes in the environment. It has been reported that a number of biological systems such as neural firing, animal motion and gene regulation [2–4, 8] indeed display behaviour evidencing their operation near criticality. Such critical behaviour is also present in ecological systems, which can therefore be prone to small perturbations, and are commonly said to reside near a tipping point. It is commonly believed that there are three modes of operation which are shared by all critical biological systems: (i) stable, (ii) critical, (iii) chaotic (super-critical). In the seminal work of Kaufman [1, 5] it was first stated that gene regulatory networks (GRNs) are critical and (random) Boolean networks have been developed to study their critical behaviour. An explanation behind why biological systems are poised at criticality, however, is still lacking [2, 8]. In this paper we address this question by developing a new nonlinear minimal model of interacting co-evolving GRNs to help shed light on this intriguing question. In contrast to the existing approaches, which mainly use Random Boolean Networks, we use ordinary differential equations with nonlinear boundary conditions to model the GRNs, which was originally put forward by Stokic, Hanel and Thurner [6]. Our main idea is to use similarities between critical systems in physics and in biological systems. To be able to address biological systems using techniques from statistical physics, a generalized non-equilibrium statistical mechanics [5] is needed in order to describe the properties of ensembles of complex systems with very many dynamically coupled elements. Understanding the characteristic structure and behaviour of the members of the ensemble will help to understand both the emergence of order in organisms and its adaptive evolution.

2 Results

We start from an existing model for a single regulatory network with dynamics as proposed by Stokic, Hanel and Thurner [6, 7], which we generalize to accommodate many co-evolving GRNs. The set of linear stochastic evolution equations is complemented

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

with nonlinear constraints ($x_i^\alpha > 0$). The concentrations of mRNA (x_i^α) evolve according to

$$\frac{dx_i^\alpha}{dt} = \sum_j A_{ij}^\alpha x_j + J_i^\alpha + r_i^\alpha \tag{1}$$

Here A_{ij}^α is the weighted adjacency matrix of the full (autocatalytic) reaction network, whose entries may be zero, positive or negative, indicating that species i is stimulated, not affected or supressed by species j. J_i^α models the flow of molecular species i into (> 0) or out of (< 0) the system. Typically J_i^α arises as a consequence of the fact that the average concentration x_i^α is positive. The noise terms r_i comprise both multiplicative and additive noise. For a single network this model can generate oscillatory, chaotic and stable behaviour [6]. In particular it can be numerically calculated for what kind of network structure and in particular for which average degree k ($10 < k < 25$ in [6]) the network is critical by computing the largest Lyapunov exponent of the equations (1). Largest Lyapunov exponents near zero signal critical behaviour.

Here we also take evolution of the networks A_{ij}^α into account. How the network topology will evolve in time is an extremely relevant question in biology, but has so far been hardly addressed. In this paper we discuss different classes of evolution. One important class of evolution equations may be obtained by using the so-called Kullback-Leibler divergence between the rows of matrix A_{ij}^α and A_{ij}^β ($\mathrm{D_{KL}}(A_{ij}^\alpha | A_{ij}^\beta)$) where A_{ij}^β constitutes an average of all GRNs, which quantifies the information loss when the matrix A_{ij}^α is used to estimate the environment constituted by A_{ij}^β. By requiring the ($\mathrm{D_{KL}}(A_{ij}^\alpha | A_{ij}^\beta)$) to be minimal, we can evolve the adjacency matrices in time. We numerically show that this indeed leads to a critical state, that is, the evolution naturally gives rise to networks with an average degree in the range of the critical networks; see Fig. 1

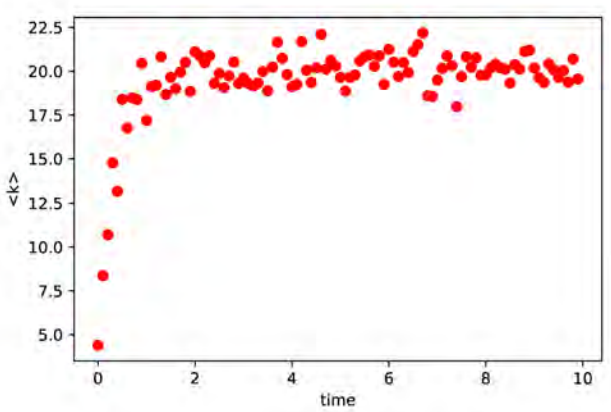


Fig. 1. The behaviour of the average value of the degree $< k >$ as a function of the time t. The network ends in the range where the system is critical, that is $k \in [10, 25]$.

Summary. We have studied a dynamical system on a network that models GRNs. The nonlinearities in the system give rise to interesting behaviour depending on the network topology. The topology that arises when the network evolves according to a principle of minimal Kullback-Leiber divergence is such that indeed the network evolves to a state where the Lyapunov exponent is nearly zero.

References

1. S. Kaufman, J. Theor. Biol. **119**, 1 (1986).
2. M. Munoz, Criticality and dynamical canaling in living systems, Rev. Mod. Phys. **90**, 031001, (2018).
3. J.M. Beggs, The criticality hypothesis: How real cortical networks might optimize information processing, Phil. Trans. R. Soc. A, **366**, 329 (2008).
4. A. Kamimura and K. Kaneko, Phys. Rev. Lett. **105**, 168103 (2010).
5. S. Kaufman, The origins of order, self organization and selection in evolution, OUP 1993.
6. D. Stokić, R. Hanel, S. Thurner, Inflation at the edge of chaos in a simple model of gene interaction networks, Phys. Rev E. **77**, 061917 (2008).
7. R. Hanel, M. Pöchacker, S. Thurner, Living on the edge of chaos: minimally nonlinear models of genetic regulatory dyanmics, Phil. Trans. R. Soc. A **368**, 5583 (2010).
8. Th. Mora, W. Bialek, Are Biological systems poised at criticality?, J. stat. Phys. **144**, 268 (2011).

Understanding the Primary-Specialty Referral Mechanism using Network Science

João Casal da Veiga[1], Qiwei Han[1], and Claudia Soares[2]

[1] Nova School of Business and Economics, Carcavelos, Portugal
[2] Instituto Superior Técnico, Lisbon, Portugal

1 Introduction

The medical referral between Primary Care Physicians (PCP) and specialists represents the formal mechanism in the health system to address the need of patients for specialty care [5]. Typically, for a given patient with clinical needs, PCP can make a choice of several specialists to whom they may refer and their choice would have important downstream effects. As such, primary-specialty referral may affect many aspects of patient care, such as quality of care, patient satisfaction and health care costs, etc. [3].

Researchers recently leveraged the patient consultation history extracted from insurance claims data to construct the patient sharing network between physicians based on the shared patients [3, 6]. Essentially, patient sharing network operationalizes an informal information-sharing network in which physicians provide care to shared patients. This network does not necessarily conform to the formal organizational structure that physicians are affiliated with, but may provide valuable insights in explaining the referral mechanism. For example, both [3] and [7] performed social network analysis on Medicare administrative data and showed that structure of patient sharing networks and the position of physicians in the network has a significant relationship with the overall cost and intensity of care. [2] further discovered small-world structure and strong correlations between certain network statistics with health system statistics. These metrics derived from network science can serve as informative features to boost predictive model performance and optimize health system for improved medical outcomes [1].

2 Primary-Specialty Referral Network Analysis

In this paper, we aim to add to the literature of understanding the primary-specialty referral mechanism. We obtain a large-scale patient consultation dataset from a private European health provider with over 9 million consultations between 1.3 million patients and 2,308 physicians (515 PCP and 1793 specialists) in 7 hospitals between 2012-2017. The primary-specialty referral is defined as when a patient consults a PCP and then a specialist within 30 days. In other words, there only exist links between two distinct set of physician nodes, namely PCP and specialists. As such, we develop a weighted bipartite network where 460 PCP are connected with 1,542 specialists through 78,593 edges. The edge weight of referral network represents the number of patients that PCP refer to the specialists. Importantly, 306 physicians do not have any edge to the referral

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

network, which raises the potential inefficiency concerns for their lack of involvement in referral process.

Besides, we obtain additional physician registration data from the Human Resource department of the provider, including gender, age, education and internship institution, specialty, working hospital, etc. As we augment the referral network with such information, summary statistics show PCP and specialists indeed have different background. The gender composition for PCP is 69% of female and 31% male, and 43% of female and 57% of male for specialists. Also, we observe certain level of homophily. For example, more than half of the referrals are made between the physicians with the same gender and more than 70% of the referrals are made between those with the age difference less than 10 years. More interestingly, 67% of the referrals are made between physicians that work at the same hospital, implying that referral network may exhibit strong clustering in terms of physicians' background.

Following [8], we compute macro-level structural metrics from the resulting referral network and compare them against those of a synthetic Erdős-Rényi random network. We find that degree distributions (the number of specialists that PCP refers to, and the number of PCPs that specialist receives from) for the referral network do not follow Poisson distribution. Meanwhile, we obtain average clustering coefficient for the referral network (0.17) to measure the fraction of the number of observed squares to the total number of possible squares in the network. It is about 2.5 times higher than that of random network (0.07). This represents an essential precondition for referral network to exhibit small-world structure and suggests that physicians in the referral network have higher tendency to cluster together. We also quantify both betweenness and closeness centrality for physicians in the referral network. The former describes the number of shortest paths that pass through a physician while the latter describes the reciprocal of the sum of distance to all other physicians in the network. Top 25% and bottom 25% of PCP in terms of betweenness centrality initiate 58.4% and 0.07% of referrals, respectively. Top 25% and bottom 25% of specialists in terms of betweenness centrality receive 55.9% and 3.3% of referrals, respectively. Again, this raises the inefficiency concern as referral process occurs high skewed towards a small number of physicians.

We adopt the popular modularity-based optimization algorithm Louvain to extract communities from the referral network [4]. In total, we identify 7 distinct communities, which happens to correspond to the number of hospitals of the provider. Figure 1 shows the visualization of community structure. In general, PCP tend to refer patients to specialists belonging to the same community, which indicates that physicians may form a "referral clique" wherein referral process occurs more likely than to physicians from different communities. Meanwhile, there is one community (in purple) that is located distantly from other communities, which contains physicians mostly working at the hospital in a different region. Moreover, we demonstrate that physicians within the same community share more similarity in terms of their background, namely, they are at the similar age, have similar number of years of experience, used to study and intern at the same institution and now work at the same hospital. Our results show that referral network may highly overlap with the social network of physicians and in the future work we plan to explore the correlation between them.

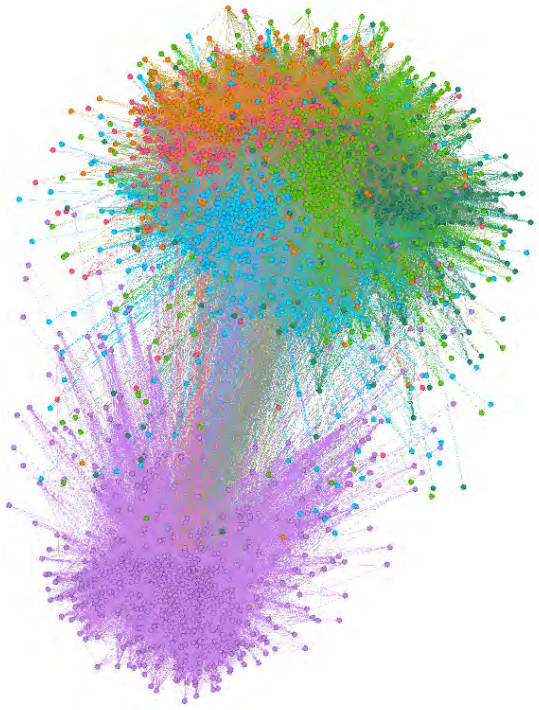


Fig. 1. Communities extracted from the referral network

References

1. An, C., O'Malley, A.J., Rockmore, D.N.: Referral paths in the U.S. physician network. Applied Network Science **3**(1) (2018)
2. An, C., O'Malley, A.J., Rockmore, D.N., et al.: Analysis of the U.S. Patient Referral Network. Statistics in medicine **37**(5) (2018) 847–866
3. Barnett, M.L., Christakis, N.A., O'Malley, A.J., et al.: Physician Patient-Sharing Networks and the Cost and Intensity of Care in US Hospitals. Medical Care **50**(2) (2012) 152–160
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., et al.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008)
5. Forrest, C.B., Nutting, P.A., Von Schrader, S., Rohde, C., Starfield, B.: Primary care physician specialty referral decision making: Patient, physician, and health care system determinants. Medical Decision Making **26**(1) (2006) 76–85
6. Landon, B.E., Keating, N.L., Barnett, M.L., et al.: Variation in patient-sharing networks of physicians across the United States. JAMA: Journal of the American Medical Association **308**(3) (2012) 305–311
7. Landon, B.E., Keating, N.L., Onnela, J.P., Zaslavsky, A.M., Christakis, N.A., James O'Malley, A.: Patient-sharing networks of physicians and health care utilization and spending among medicare beneficiaries. JAMA Internal Medicine **178**(1) (2018) 66–73
8. Latapy, M., Magnien, C., Vecchio, N.D.: Basic notions for the analysis of large two-mode networks. Social Networks **30**(1) (2008) 31–48

Alternative mRNA Splicing-based Drug Response Networks Yield Interactive and Mechanistic Insights

Edward Simpson[1,2], Jill Reiter[2], and Yunlong Liu[1,2]

[1] Indiana University-Purdue University, Indianapolis IN 46202, USA
[2] Indiana University School of Medicine, Indianapolis IN 46202, USA
edrsimps@iupui.edu

1 Introduction

Alternative splicing is a tightly-controlled transcriptional regulatory mechanism where exons can be selectively included or excluded during pre-mRNA processing. These exons may have essential roles in protein structure and function. Splicing is composed of a variety of event type definitions which are characterized by exon positions and rules governing exon usage. Skipped-exon (SE) events are the most common and refer to the inclusion or exclusion of single exons. Splicing has a significant influence on many aspects of cellular physiology, including cellular identity, plasticity, signaling and cancer [1,2]. Splicing has also been linked to cancer drug resistance [3,4]. Certain spliced isoforms can manipulate kinase signaling and alter cellular drug response [5,6]. Despite this, few studies have explored connections between drug response and splicing. It is well known that many drugs exploit similar targets or pathways as development of structurally homologous compounds is cheaper and faster than development of novel therapeutics. Yet, to the best of our knowledge, no one has investigated the commonality between predictive splicing signatures and various related or unrelated compounds.

In previous work we identified differentially spliced SEs in pre-treatment transcriptional profiles from cancer cell lines, the SEs' relationships with drug response and the regulatory elements that play a role in their splicing [7]. We found that alternatively spliced SEs were highly predictive of doxorubicin drug response. Additionally, extending the same modeling approach to other drugs yielded similar results. We then hypothesized that drugs from the same class or with similar activities would share predictive splicing features. Here, we expand our work to incorporate other categories of alternatively spliced events and construct tissue-specific drug networks utilizing common predictive splicing features. We describe the drug network characteristics and explore individual drug modules across networks.

2 Methodology

RNAseq data for 975 cell lines from the Cancer Cell Line Encyclopedia (CCLE) were integrated with drug-response data for 501 drugs (tested in 860 cell lines) from the Cancer Therapeutic Response Portal (CTRP) [8,9]. The number of cell lines with both RNAseq and drug response data differed by drug. RNAseq data was mapped with STAR using Hg19 and GRCh37v87 annotation [10,11]. A list of spliced events in reference

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

genes was generated from the annotation GTF file and junction read counts for each splice variant were collected. For each drug, cell lines were separated into sensitive and resistant groups based on the quantile of the area under the concentration-response curve (AUC) value in CTRP; 33% or less were considered sensitive and 66% or above were considered resistant. Differentially-spliced events across sensitive and resistant cell lines were identified for each drug by using a previously published quasi-binomial generalized linear modeling framework and applying a FDR cutoff of <= 0.01 [7]. To select for validity and biological significance, we required each event to have at least one splice-junction read in >= 35% of all cell lines and a minimum difference in the mean fraction of inclusive to total junction reads between sensitive and resistant cell lines >= 0.10. Drug-drug networks for each tissue were constructed using a modified Jaccard index for edge weight:

$$W_{ab} = \frac{(A \cap B) - D_{ab}}{A \cup B} \tag{1}$$

where D_{ab} = divergent, (i.e. the number of exons observed to have a higher inclusion level in sensitive cells of drug A but lower inclusion level in sensitive cells of drug B). Module identification was accomplished in three steps. First, hierarchical clustering was performed on the network matrix using average distance. Next, all clusters in the bottom 15% of tree heights with between 3 and 15 members were extracted and merged if one contained all members of another. Finally, clusters were filtered for significance <=0.05 using their b- and c-scores, which are metrics analogous to the probability of observing a module in the random network given the network size, module size, inner- and outer-degrees [12]. Modules were annotated with drug activity from CTRP. Differentially-spliced events present across multiple drugs in a module were identified and annotated with gene symbol, protein structure, function and domain information.

3 Results

We combined two large public datasets, CCLE and CTRP, to maximize the number of cell lines and different classes of drugs in the networks. Cell lines were grouped by tissue type and the two groups containing the most cell lines, haematopoietic & lymphoid (HL) and lung, were selected for further study. We observed a total of 437 connected drugs with 38,802 edges in the HL tissue network and 441 connected drugs with 43,371 edges in the Lung tissue network (**Table 1**). Both networks exhibited random network

Table 1. Tissue-specific Network Summary

Tissue	Nodes	Edges	Unfiltered Modules	Filtered Modules
HL	437	38,802	35	14
Lung	441	43,371	30	11

structure and upon bi-partite inspection, there appeared to be a small number of drugs with many differentially spliced events. These drugs created hairballs in the network by facilitating weak connections with many smaller degree nodes and made module

identification more difficult. Therefore, we applied a module identification strategy that would take advantage of the expected behavior of drugs in a network while minimizing the size of extracted modules. Some of the stronger modules we identified, as defined by larger size and a high average W_{ab}, shared many of the same drug members in both tissue types. These modules primarily contained chemotherapeutic agents such as the platinum and anthracycline families of anti-cancer drugs. Other modules were more specific to the respective tissue and members clustered with different partners. One such example is erlotinib, an EGFR tyrosine kinase inhibitor (see **Fig. 1**). Erlotinib modules from HL and Lung networks both included tyrosine kinase and EGFR inhibitors.

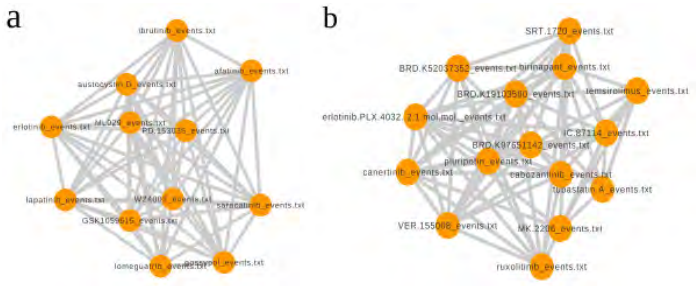


Fig. 1. Erlotinib modules from HL and lung networks are highly connected and share similar activity but different compound identities. **a.** HL module, out of 513 events 70 were found significant in more than one compound. **b.** Lung module, out of 332 events, 69 were present in more than one compound.

4 Conclusions

We found network analysis using pre-treatment splicing information shows drugs of similar activity cluster together by having common splicing features associated with drug response; however, clusters may not retain the same cluster partners or predictive splicing features in other tissues. We suspect this is due to tissue-specific splicing regulation and that drugs of the same class may have altered activity in certain tissues. We intend to further characterize this and other observations from the networks.

References

1. Baralle, F.E., Giudice, J.: Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. 18, 437 (2017)
2. David, C.J., Chen, M., Assanah, M., et al.: HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature. 463, 364–368 (2010)
3. Dehm, S.M.: mRNA Splicing Variants: Exploiting Modularity to Outwit Cancer Therapy. Cancer Res. 73, 5309–5314 (2013)
4. Zammarchi, F., Stanchina, E. de, Bournazou, E., et al.: Antitumorigenic potential of STAT3 alternative splicing modulation. Proc. Natl. Acad. Sci. 108, 17779–17784 (2011)

5. Cesi, G., Philippidou, D., Kozar, I., et al.: A new ALK isoform transported by extracellular vesicles confers drug resistance to melanoma cells. Mol. Cancer. 17, (2018)
6. Peng, H., Peng, T., Wen, J., et al.: Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach. Bioinformatics. 30, 1899–1907 (2014)
7. Simpson, E., Chen, S., Reiter, J.L., Liu, Y.: Differential Splicing of Skipped-Exons Predicts Drug Response in Cancer Cell Lines. Genomics Proteomics Bioinformatics. [In Press]
8. Barretina, J., Caponigro, G., Stransky, N., et al.: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 483, 603–607 (2012)
9. Basu, A., Bodycombe, N.E., Cheah, J.H., et al.: An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 154, 1151–1161 (2013)
10. Dobin, A., Davis, C.A., Schlesinger, F., et al.: STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29, 15–21 (2013)
11. Zerbino, D.R., Achuthan, P., Akanni, W., et al.: Ensembl 2018. Nucleic Acids Res. 46, D754–D761 (2018)
12. Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S.: Finding Statistically Significant Communities in Networks. PLOS ONE. 6, e18961 (2011)

Part II

Community Structure

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Mean Consensus Time of the Voter Model on Networks Partitioned into Two Cliques of Arbitrary Sizes

Michael T. Gastner and Kota Ishida

Yale-NUS College, Division of Science, 16 College Avenue West, #01-220 Singapore 138527
michael.gastner@yale-nus.edu.sg,
WWW home page: http://michaelgastner.com/

1 Introduction

The voter model is a paradigmatic agent-based model that represents opinion dynamics in social networks. The dynamics of the model consists of repeatedly choosing one agent uniformly at random. The selected agent then copies the current opinion of a randomly selected neighbour. As long as the network is connected and finite, this update rule guarantees that the agents must eventually reach a consensus after a finite time T. The mean consensus time $\langle T \rangle$ depends on the initial distribution of opinions and the network structure.

While early studies of the voter model focused on complete graphs or regular lattices, interest has recently shifted towards networks with more complex topologies, for example networks with a community structure [1], [2], [3]. Here we analyze the voter model on the simplest possible multi-community network: two cliques (i.e. fully connected subgraphs) connected by a small number X of intercommunity edges (Figure 1). Previous work on networks with two equally large cliques has shown that the mean consensus time $\langle T \rangle$ is proportional to the number N of vertices in the network unless the connections between the cliques are extremely sparse [2]. Because $\langle T \rangle \propto N$ is the same scaling relation as in the case of a single-clique network [4], it has been argued that community structure is of limited importance for the voter model. Here we show that, on the contrary, the two-clique topology gives rise to many intriguing features.

2 Results

Let us denote by α the relative fraction of vertices in clique 1. For example, in the network depicted in Figure 1, α is equal to $\frac{7}{12}$. For all values of α, sparsely connected cliques need a long time to reach a consensus, as one might intuitively expect. Counterintuitively, however, additional links between the cliques do not necessarily speed up the consensus (except in the special case $\alpha = \frac{1}{2}$). Instead, numerical simulations (Figure 2) show that there is an optimal intermediate connectivity that minimizes $\langle T \rangle$. The simulations suggest that the optimal number of interclique edges scales as $X_{\min} \propto N^{3/2}$, which puts the optimum between the case of a constant number of interclique edges per agent ($X_{\min} \propto N$) and a complete graph ($X_{\min} \propto N^2$). Hence, to accelerate a consensus between cliques, agents should reach out to members in the other clique, but not to the extent that cliques lose their identity as distinct communities.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

We confirm the numerical results with an equation-based analysis. For the sake of simplicity, we show the equations only for the case of a polarized initial condition (i.e. both cliques are internally unanimous, but there is disagreement between the cliques). Similar results can be derived for other initial conditions. We make two heterogeneous mean-field approximations for the consensus time $\langle T \rangle$:

– a Taylor expansion for small X,

$$\langle T \rangle \approx t_{\text{sparse}} = \frac{\alpha^2(1-\alpha)^2 N}{X d(\alpha,N,X)} \big[2(2\alpha^2-2\alpha+1)X^3 + 2(\alpha^2-\alpha+1)NX^2 + \alpha(1-\alpha)(2\alpha^2-2\alpha+3)N^2X + \alpha^2(1-\alpha)^2N^3\big] \tag{1}$$

with the auxiliary function

$$\begin{aligned} d(\alpha,N,X) =& (3\alpha^2-3\alpha+1)(2\alpha^2-2\alpha+1)X^2 \\ &+ \alpha(1-\alpha)(4\alpha^4-8\alpha^3+11\alpha^2-7\alpha+2)NX \\ &+ \alpha^2(1-\alpha)^2(2\alpha^2-2\alpha+1)N^2, \end{aligned} \tag{2}$$

– an adiabatic approximation for large X,

$$\langle T \rangle \approx t_{\text{dense}} = - \frac{\alpha(1-\alpha)N[(2\alpha^2-2\alpha+1)N^2+2X]^2}{\alpha(1-\alpha)N^2[(3\alpha^2-3\alpha+1)N^2+2X]+X^2} [m\ln m + (1-m)\ln(1-m)], \tag{3}$$

where

$$m = \frac{(\alpha^2N^2-\alpha N+X)}{(2\alpha^2-2\alpha+1)N^2-N+2X}\,. \tag{4}$$

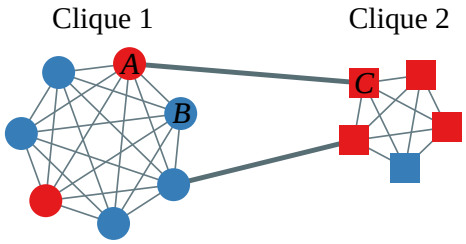


Fig. 1. Small illustrative example of a two-clique network. Each vertex represents an agent that has exactly one of two possible opinions: "red" or "blue". In this example, clique 1 is a complete graph with 7 vertices, whereas clique 2 has only 5 vertices. The cliques are connected by two intercommunity edges (thick lines). In our analysis, we vary the relative sizes of the two communities and the number of intercommunity edges. We apply the update rules of the voter model. That is, we first choose a random focal vertex, for example A in the depicted network. Then we choose a random neighbour of the focal vertex and copy the neighbour's opinion. In our example, if the chosen neighbour is B, A changes its opinion to blue. However, if the chosen neighbour is C, A keeps its current (i.e. red) opinion.

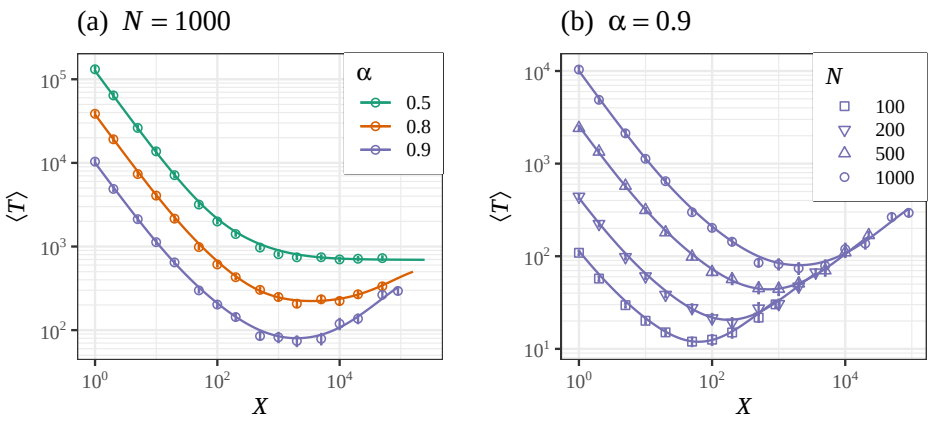


Fig. 2. Mean consensus time $\langle T \rangle$ as a function of the number of interclique edges X. Point symbols represent simulation results. In all simulations, the initial opinions are completely polarized: both cliques are internally unanimous, but there is disagreement between the cliques. The curves are equation-based predictions. In (a), we fix the number of vertices to be $N = 1000$ and vary the fraction α of vertices in the first clique. If $\alpha \neq \frac{1}{2}$, the minimum of $\langle T \rangle$ is attained at an intermediate value of X, where the cliques are neither sparsely nor fully connected. In (b), we fix $\alpha = 0.9$ and vary N. The value $X_{\min}$ that minimizes $\langle T \rangle$ is proportional to $N^{3/2}$.

By interpolating between the two asymptotic approximations, we obtain an equation for $\langle T \rangle$ that is in excellent agreement with the simulations for all values of X,

$$\langle T \rangle = t_{\text{dense}}(X) + t_{\text{sparse}}(X) - \lim_{X' \to \infty} t_{\text{sparse}}(X'). \tag{5}$$

This interpolation is shown by the curves in Figure 2. From equations (1)–(5) it can be shown that $X_{\min} \propto N^{3/2}$ [5], consistent with the numerical results.

Summary. We show that, counterintuitively, the mean consensus time $\langle T \rangle$ is typically not a monotonically decreasing function of interclique connectivity. To minimize $\langle T \rangle$, the optimum number of interclique edges $X_{\min}$ should scale as $X_{\min} \propto N^{3/2}$, where N is the number of vertices. Consequently, to reach a consensus quickly, the agents must strike a balance between a sparse and a dense interclique connectivity.

References

1. Castelló, X., Toivonen, R., Eguíluz, V.M., Saramäki, J., Kaski, K., Miguel M.S.: Anomalous lifetime distributions and topological traps in ordering dynamics. EPL 79(6), 66006 (2007)
2. Masuda, N.: Voter model on the two-clique graph. Phys. Rev. E 90(1), 012802 (2014)
3. Bhat, D., Redner, S.: Opinion Formation under Antagonistic Influences. arXiv:1907.13103 (2019)
4. Sood, V., Antal, T., Redner, S.: Voter models on heterogeneous networks. Phys. Rev. E 77(4), 041121 (2008)
5. Gastner, M.T., Ishida, K.: Voter model on networks partitioned into two cliques of arbitrary sizes. arXiv:1908.00849 (2019)

Clustering via Hypergraph Modularity

Bogumił Kamiński[1], Bartosz Pankratz[1,3], Valérie Poulin[2], Paweł Prałat[3], Przemysław Szufel[1], and François Théberge[2]

[1] SGH Warsaw School of Economics, Warsaw, Poland
[2] The Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada
[3] Department of Mathematics, Ryerson University, Toronto, ON, Canada

1 Introduction

Despite the fact that many important problems (including clustering) can be described using hypergraphs, theoretical foundations as well as practical algorithms using hypergraphs are not well developed yet. In [2], we proposed a hypergraph modularity function that generalizes its well established and widely used graph counterpart measure of how clustered a network is. In order to define it properly, we generalized the Chung-Lu model for graphs to hypergraphs. Moreover, some theoretical foundations about our hypergraph modularity function as well as some simple experiments on synthetic hypergraphs are provided. In particular, we showed that a strict version of our proposed modularity function often leads to a solution where a smaller number of hyperedges gets cut as compared to optimizing modularity of 2-section graph of a hypergraph. The conclusion is that the proposed novel approach to deal with hypergraphs yields substantially different clusters than its 2-section graph counterpart. It is different but the question is: is it better or worse?

In order to answer this question, we work on developing fast algorithms for clustering on hypergraphs. We have implemented a `SimpleHypergraphs.jl` library[4] using the Julia language [1]. In this way our algorithms are computationally efficient and easy to develop and maintain at the same time. Our next step is to perform more experiments on real networks that are naturally represented as hypergraphs, see Section 3.

The presented research was partially financed by NAWA — The Polish National Agency for Academic Exchange.

2 Theoretical Foundations

Review of Graph Modularity. The definition of modularity for graphs was first introduced by Newman and Girvan in [3].

For a graph $G = (V,E)$ and a given partition $\mathbf{A} = \{A_1, \ldots, A_k\}$ of V, the modularity function is defined as follows:

$$q_G(\mathbf{A}) = \sum_{A_i \in \mathbf{A}} \left(\frac{e_G(A_i)}{|E|} - \frac{(vol(A_i))^2}{4|E|^2} \right), \tag{1}$$

[4] `https://github.com/pszufe/SimpleHypergraphs.jl`

where $e_G(A_i) = |\{\{v_j, v_k\} \in E : v_j, v_k \in A_i\}|$ is the number of edges in the subgraph of G induced by the set A_i, and $vol(A_i) = \sum_{v_j \in A_i} deg(v_j)$. The modularity measures the deviation of the number of edges of G that lie inside parts of $\mathbf{A}$ from the corresponding expected value based on the Chung-Lu distribution $\mathscr{G}(G)$ [4]. The *modularity* $q^*(G)$ is defined as the maximum of $q_G(\mathbf{A})$ over all possible partitions $\mathbf{A}$ of V.

Hypergraph Modularities. In [2], we generalized the Chung-Lu model to hypergraphs and used it as a null model allowing us to define hypergraph modularity. Consider a hypergraph $H = (V, E)$ and $\mathbf{A} = \{A_1, \ldots, A_k\}$, a partition of V. For edges of size greater than 2, several definitions can be used to quantify the edge contribution given $\mathbf{A}$, e.g.:

(a) all vertices of an edge have to belong to one of the parts to contribute; this is a *strict* definition that we focus on in this paper;
(b) the *majority* of vertices of an edge belong to one of the parts;
(c) at least 2 vertices of an edge belong to the same part; this is implicitly used when we replace a hypergraph with its 2-section graph representation.

We see that the choice of hypergraph modularity function is not unique and it depends on how strongly we believe that a hyperedge is an indicator that vertices belonging to it fall into one community.

In [2], we derived a general formula that covers all variants but here, for illustration purpose, we concentrate only on the extreme case, option (a), that we call *strict*. The strict modularity function of a hypergraph partition $\mathbf{A}$ is defined as follows:

$$q_H(\mathbf{A}) = \frac{1}{|E|}\left(\sum_{A_i \in \mathbf{A}} e(A_i) - \sum_{d \geq 2} |E_d| \sum_{A_i \in \mathbf{A}} \left(\frac{vol(A_i)}{vol(V)}\right)^d\right), \tag{2}$$

where $E_d \subseteq E$ is the set of hyperedges of size d and $vol(V) = \sum_{d \geq 2} d \cdot |E_d|$. Just as for graphs, the corresponding *modularity* q_H^* is defined as the maximum of $q_H(\mathbf{A})$ over all possible partitions $\mathbf{A}$ of V.

Note that similarly to graphs, finding a graph partition that yields the highest modularity is an NP-hard problem. Within the fore-mentioned SimpleHypergraphs.jl library we are working on heuristics for detection of communities which are also discussed in [2].

3 Experiments

In [5] we performed some initial experiments on a hypergraph obtained from Yelp data-set which consists of thousands of nodes (restaurants) millions of their reviews (that from hyperedges). Additionally, we observed that the additional information conveyed via hypergraphs (as opposed to their 2-section representations) lead to better partitioning of the vertices in the analyzed data-set with respect to considered ground truth.

Now we want to test the hypergraph-based approach in web-graph applications. The growth of Internet usage in last two decades has created unprecedented opportunities for social scientists. Digital services, especially social media, are amazing reservoir of data, holding valuable insights about social systems. As a result, researchers are able to

experiment on real systems on the previously unprecedented scale. The usage of the social media data allows to better understand the dynamics of the protests movements [6, 7] and polarisation of the political debate [8]. However, Internet also has changed the way how the people behave and communicate. The rise of the social media creates new mechanisms shaping the society.

In this work we show that hypergraphs, are superior to their traditional counterparts which are widely used in the network science. In the context of the social network, the representation of the social circles (e.g. followers on Twitter or friends on Facebook) as a hyperedges seems more natural than the edges connecting only two nodes.

In order to prove that, we design an experiment measuring the polarization of political views of Twitter users. The phenomenon of the Internet bubbles or echo chambers (closed groups of users showing strong resemblance and interacting mostly inside specific clusters) is crucial to understand the way how the political debate is or might be shaped by the different actors. The selective exposure to sources of information makes the citizens more prone to the discourse framing techniques such as fake news [9] or political bots [10].

By using tweets concerning different political and nonpolitical issues we build hypergraphs and then measure the strength of their community division and similarities between nodes in each cluster. As a result, we are able to detect the most important topics and better understand the online communication patterns regarding different subjects. Finally, we compare obtained results to the clustering of the regular graphs build around the same data and results previously obtained in the literature.

References

1. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: A fresh approach to numerical computing. SIAM review 59(1), 65–98 (2017).
2. B. Kaminski, V. Poulin, P. Pralat, P. Szufel, and F. Theberge, Clustering via Hypergraph Modularity, preprint, arXiv:1810.04816.
3. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys. Rev. E. 2004; 69: 026–113.
4. Chung FRK, Lu L. Complex Graphs and Networks. American Mathematical Society; 2006.
5. A. Antelmi, G. Cordasco, B. Kaminski, P. Pralat, V. Scarano, C. Spagnuolo, and P. Szufel, SimpleHypergraphs.jl — Novel Software Framework for Modelling and Analysis of Hypergraphs, Proceedings of the 16th Workshop on Algorithms and Models for the Web Graph (WAW 2019), Lecture Notes in Computer Science 11631, Springer, 2019, 115–129.
6. Barber P, Wang N, Bonneau R, Jost JT, Nagler J and Tucker J. The Critical Periphery in the Growth of Social Protests. PLoS ONE 10(11) (2015).
7. Gallagher RJ, Reagan AJ, Danforth CM and Dodds PS. Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. PLoS ONE 13(4) (2018).
8. Barber P, Jost JT, Nagler J, Tucker J and Bonneau R. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? Psychological Science 26(10), 1531–1542 (2015).
9. Allcott H, Gentzkow M. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives 31(2), 211–236 (2017).
10. Woolley SC, Guilbeault DR. Computational propaganda in the United States of America: Manufacturing consensus online. Computational Propaganda Research Project (2017).

Community Detection with Eigenvector and Katz Centrality

Mark Ditsworth[1] and Justin Ruths[1]

University of Texas at Dallas, Richardson TX 75080, USA,
markditsworth@protonmail.com,
WWW home page: http://justinruths.com

1 Introduction

The computational demands of community detection algorithms such as Louvain and spectral optimization can be prohibitive for large networks. These community detection algorithms require either numerous iterations through combinatorial partitions of the network nodes or linear algebraic operations on the adjacency matrix. More recent literature on community detection extends the use of these methods, altering calculations and processes, but do not deviate from the iterative maximization of modularity. For large complex networks the computational and memory requirements often prove impractical.

This work demonstrates the utility of Katz centrality and eigenvector centrality as indicators of community membership in large undirected networks. This method is shown to produce well-defined communities (when sufficient modularity is present in the network) in a much faster runtime than Louvain. Based on our datasets our proposed approach has runtimes as low as 8.6% of the Louvain community detection runtime for smaller networks, and 0.002% of the Louvain runtime for larger networks.

2 Methods

Eigenvector centrality is based on the idea that a node's importance is related to the importance of its neighbors. The eigenvector centrality of node i (x_i) is measured by the scaled sum of the eigenvector centralities of its neighbors,

$$\mathbf{x} = \frac{1}{\lambda_1}\mathbf{A}\mathbf{x}, \tag{1}$$

where λ_1 is the leading eigenvalue of the adjacency matrix $\mathbf{A}$, and $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ [1]. This is clearly the equation for the leading eigenvector of the adjacency matrix, thus it only describes the most dominant mode of the network. In modular networks, this has been shown to confine the large eigenvector centrality values to a certain collection of nodes, despite the existence of other collections that appear to have similar importance [2].

Katz centrality is calculated similar to eigenvector centrality, but with free centrality β given to all nodes and α chosen such that $\alpha < \frac{1}{\lambda_1}$ [3],

$$\mathbf{x} = (I - \alpha\mathbf{A})^{-1}\,\beta\mathbf{1}. \tag{2}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

As shown in [2], the inverse operation above can be expressed in terms of its power series, allowing Katz centrality to be equivalently written as

$$\mathbf{x} = a_1\mathbf{u}_1 \sum_{k=0}^{\infty} (\alpha\lambda_1)^k + \cdots + a_n\mathbf{u}_n \sum_{k=0}^{\infty} (\alpha\lambda_n)^k, \tag{3}$$

with real coefficients $a_1, \ldots, a_n$. Since $\alpha < \frac{1}{\lambda_1}$, each infinite sum will converge and reveals that Katz centrality spans the entire eigenbasis of $\mathbf{A}$. Thus the localization of centrality in modular networks will be significantly reduced compared to eigenvector centrality. In this work we leverage the localization of eigenvector centrality against the robustness of Katz centrality in sufficiently modular networks to identify the communities that give rise to the modularity by plotting them against each other (Fig. 1). Because of the radial structure of these plots, we use an algorithm similar to the radon line detection algorithm [4] to perform cluster identification in the Katz vs Eigenvector Centrality (KE) plot.

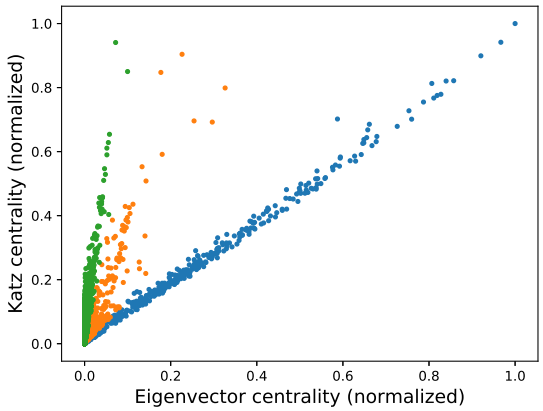


Fig. 1. Three communities discovered in the Amazon beauty review network [5].

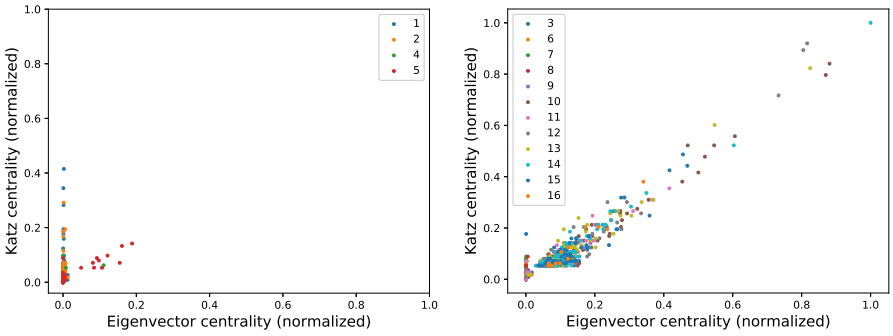
3 Results

We apply the KE community detection method on a suite of synthetic and real-world networks [5], measuring the resulting modularity (Q) of the detected communities, comparing to the results obtained from the Louvain method (summarized in Table 1). The KE method is shown to be far superior to Louvain in runtime, and to generally produce comparable modularity values indicative of high performing community detection. The resulting Q from the KE method on many of the test networks is lower than the that from Louvain. But the maximum obtainable modularity (Q_{max}) given the assigned community members is also lower than Louvain, so the normalized modularity for both methods is similar. This is likely because the two methods are extracting similar communities at different scales. For example, Louvain discovers 25 and 34 communities in the Amazon Beauty and Health networks, respectively, while the KE method discovers 3 and 2. The

combinatorial, iterative nature of Louvain method is likely causing the detection of micro-level community structures; whereas the KE method's use of a network's spectral properties causes the detection of macro-level community structures. This assertion is supported by Figure 2, where the Louvain-detected communities of an ad-hoc modular network also organize into larger communites detectable by the KE method.

	Runtime			**Modularity** (Q/Q_{max})			**N**		
Network	**L**	**S**	**KE**	**L**	**S**	**KE**	**L**	**S**	**KE**
AMZN Product	371 ms	231 ms	32 ms	0.801/0.908	0.781/0.893	0.359/0.467	14	17	3
AdHoc BA 1	11.8 s	877 ms	329 ms	0.485/0.491	0.485/0.490	0.480/0.498	2	2	3
AdHoc BA 2	2.03 m	3.88 s	228 ms	0.291/0.930	0.454/0.464	0.228/0.471	17	2	2
DBLP	12.0 m	1.15 hr	751 ms	0.805/0.982	0.713/0.974	0.019/0.034	129	191	2
AdHoc BA 3	2.07 hr	4.65 m	1.97 s	0.203/0.931	0.382/0.393	0.123/0.492	18	3	2
AMZN Beauty	11.7 hr	14.5 m	11.8 s	0.499/0.840	0.566/0.735	0.365/0.645	25	4	3
AMZN Health	16.2 d	5.30 m	35.7 s	0.413/0.543	0.00/0.00	0.423/0.608	34	1	2

Table 1. Comparison of runtime, resulting modularity, and number of detected communities (**N**) between Louvain community detection (**L**), spectral community detection (**S**), and the KE plot method of extracting communities from various networks with n nodes and m edges.



(a) (b)

Fig. 2. 16 communities detected using Louvain, reduced to two groups that largely follow the pattern utilized by the KE method.

References

1. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. Social Networks, (2001)
2. Sharkey, K. J.: Localization of eigenvector centrality in networks with a cut vertex. Phys. Rev. E 99, (2019)
3. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18, (1953)
4. Radon, J.: Uber die bestimmung von funktionen durch ihre integralwerte langs gewisser mannigfaltigkeiten. Journal of Mathematical Physics 69, (1917)
5. Leskovec, J., Krevl, A.: SNAP Datasets. https://snap.stanford.edu/data, (2014)

Autoinformation in non–Markovian diffusion systems.

Mauro Faccin[1], Michael Schaub[2,3], and Jean-Charles Delvenne[1]

[1] ICTEAM, Universit catholique de Louvain, Louvain-la-Neuve, Belgium,
[2] Institute for Data, Systems, and Society, Massachusetts Institute of Technology,
[3] Department of Engineering Science, University of Oxford

Summary. We analyse the behaviour of users navigating on a web portal with tools from Information Theory. We show that the higher order Markovianity of such dynamics is a fundamental part of the system complexity and a first order Markovian approximation can include several errors. We propose a partition detection algorithm based on an information theoretic criterion, the maximization of the auto–information, which leads to a low-dimensional model that simulates the real dynamics more closely.

1 Introduction

Dynamical systems that evolve on top of networks, such as disease spreading, information diffusion, or transportation systems are nearly ubiquitous. Analysing the properties of a such dynamical systems often offers valuable insights about the relevant connection patterns present in the underlying network. Indeed, dynamical properties have been used to find community structures [1,2], to rank nodes (e.g, via random–walk centrality and pagerank [3]), or to analyse other aspects of complex networks.

While the topology of the underlying network will add to the complexity of the system behavior, part of that complexity may emerge from the dynamics themselves: in particular, if the system behaves in a non–Markovian way and the evolution of the system depends on its own history; see for example the highly cited works [4,5,6] and a recent paper on epidemics on networks [7].

In this work we demonstrate how the non-Markovianity of a dynamics on a network should be considered when analysing the system behaviour. To this end we consider the behaviour of users navigating on a web portal of a Belgian broadcasting network. We show that this dynamics is more accurately described by a non-Markovian dynamics where memories plays a fundamental role, even though our urge to understand and simplify the system often leads the researcher to model it as a simple Markov process with no memory [3].

2 Results

The entrogram [8] (see Figure 1A) is a set of information theoretical quantities defined as follows:

$$I_i = I(x_t; x_{t-i-1}, \dots | x_{t-1}, \dots, x_{t-i}) \qquad i \in \mathbb{N}_0, \tag{1}$$

where $I(\cdot;\cdot)$ is mutual information and $X = \{\dots, x_{t-1}, x_t, x_{t+1}, \dots\}$ represents the states of the dynamics at each time–step. The entrogram provides a concise characterization

of the complexity of the dynamical system. The *area* $\sum_i I_i$ of the entrogram represents the total dynamical complexity of the system [8]. The latter can be divided into a measure of predictability (I_0), and a measure of the non-Markovian memory inherent in the dynamics ($\sum_{i\geq 1} I_i$). The dynamics displays a non-negligible amount of memory (see Figure 1A), with a Markovian behaviour of at least order three.

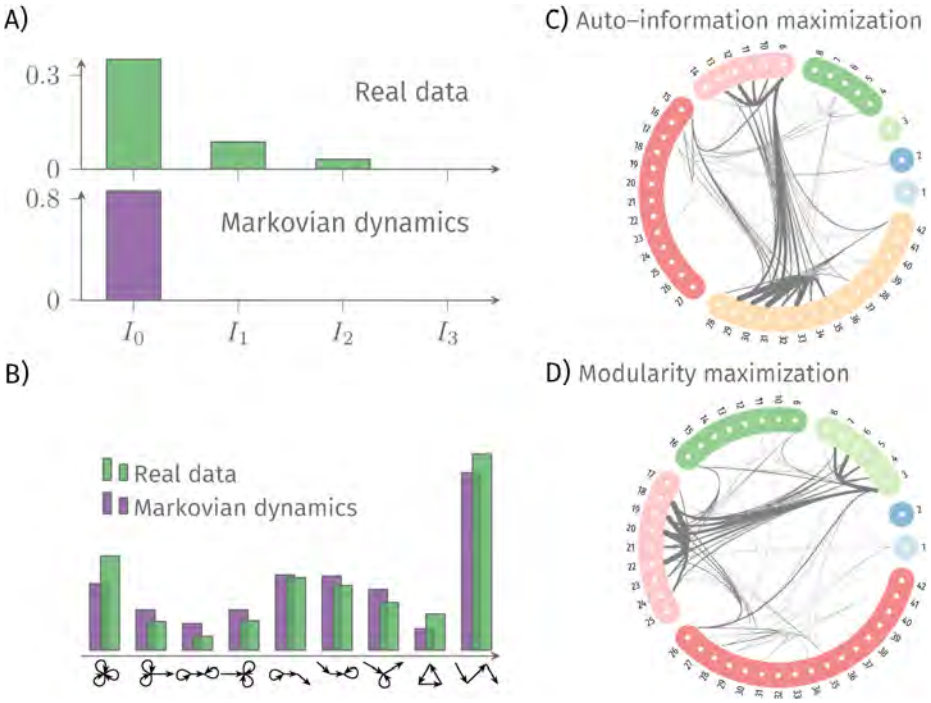


Fig. 1. Analysis of a non-Markovian dynamics of users browsing a web portal. The Entrogram [8] of the web browsing **(A)** shows clear signs of dynamical memory. The analysis of the three–steps patterns shows how two of such patterns are under-represented in a Markovian approximation of the dynamics **(B)**. Detecting communities preserving the Markovian order of the dynamics gives a slightly different partition than classical approaches **(C, D)**. We show that Markovian–order–preserving community detection can be used to better simulate the original dynamical complex system.

The analysis of the dynamical patterns contained in the dataset further supports that the real dynamics is far from being well approximated by a simple first order Markov process (see Figure 1B), and some patterns are under-represented in the latter.

As a second contribution, we introduce a state aggregation procedure that respect the Markovian order of the dynamics. To do this we maximise the auto–information between two distant time–steps in the dynamics projected to the partition space:

$$I(y_t; y_{t-T}), \tag{2}$$

where T is set to three to respect the Markov order of the original dynamics. This results in a slightly different partition of the web portal pages as compared to classical partitioning algorithms, see Figures 1C and 1D. Despite the apparent small distance between the two partitions, simulating a Markovian evolution on the reduced graph obtained from maximizing the auto–information leads to a dynamics sensibly closer to the original non–Markovian dynamics.

References

1. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences **104**(18) (2007) 7327–7331
2. Schaub, M.T., Delvenne, J.C., Yaliraki, S.N., Barahona, M.: Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. PLOS ONE **7**(2) (2012) e32210
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30**(1) (1998) 107–117
4. Scholtes, I., Wider, N., Pfitzner, R., Garas, A., Tessone, C.J., Schweitzer, F.: Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. Nature Communications **5** (September 2014) 5024
5. Delvenne, J.C., Lambiotte, R., Rocha, L.E.: Diffusion on networked systems is a question of time or structure. Nature Communications **6**(7366) (June 2015) 7366
6. Rosvall, M., Esquivel, A.V., Lancichinetti, A., West, J.D., Lambiotte, R.: Memory in network flows and its effects on spreading dynamics and community detection. Nature Communications **5**(4630) (August 2014) 4630
7. Sherborne, N., Miller, J.C., Blyuss, K.B., Kiss, I.Z.: Mean-field models for non-markovian epidemics on networks. Journal of mathematical biology **76**(3) (2018) 755–778
8. Faccin, M., Schaub, M.T., Delvenne, J.C.: Entrograms and coarse graining of dynamics on complex networks. Journal of Complex Networks **6**(5) (2018) 661–678

Nested partitions from hierarchical clustering statistical validation

Christian Bongiorno[1] and Salvatore Micciché[2] and Rosario N. Mantegna[2,3,4]

[1] Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes CentraleSupélec, Université Paris Saclay 3 rue Joliot-Curie, 91192, Gif-sur-Yvette, France,
[2] Dipartimento di Fisica e Chimica, Università di Palermo, Viale delle Scienze, Ed. 18, I-90128, Palermo, Italy
[3] Complexity Science Hub Vienna, Josefstadter Strasse 39, 1080, Vienna, Austria
[4] Computer Science Department, University College London, 66 Gower Street, WC1E 6BT, London, UK
christian.bongiorno@centralesupelec.fr

1 Introduction

Currently, the big data revolution is changing the way many disciplines perform a large amount of experimental measures and their interpretation and modeling. In fact, due to the successes of information technology revolution and the advances in robotics many scientific experiments have today an observational character rather then a design investigating a fully controlled set up. Examples are space-time records of particles dynamics, climatological monitoring of large scale regions, earthquake investigations, brain activities, gene expressions, dynamics of social and financial systems. All these types of complex systems present datasets that are genuinely multivariate and that are recorded in the presence of sources of uncertainty (modeled as noise). Their interpretation and modeling with statistically validated data mining tools require the characterization of the hierarchical sub-units present in them. A traditional unsupervised tool for the characterization of sub-units of a complex system is hierarchical clustering. In spite of the effectiveness and simplicity of this approach the extraction of a hierarchically nested partition from a hierarchical tree is still today an open problem. The most widely used approach for cluster detection used in the scientific literature is an approach originally proposed in phylogenetics and today implemented by the algorithm called Pvclust [4]. This algorithm is widely used in many disciplines and especially in genomics. It is the standard reference in the literature but present two serious limits. The first limit concerns computational time and scalability with system size. The algorithm is relatively slow and has a limited scalability and therefore it is not appropriate for very large datasets. The second limit (partly overlapping with the previous one) is related to the open problem of how to deal with the so-called familywise error. This type of error is a source of statistical errors occurring when a large number of statistical tests is performed in parallel in a system. This type of errors originates naturally in very large datasets.

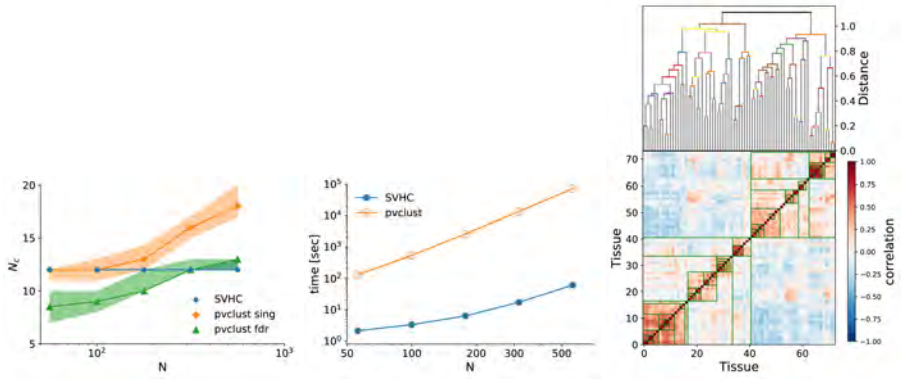
COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

In this work, we propose a greedy algorithm based on bootstrap resampling that associates a p-value at each clade of a hierarchical tree. Our algorithm gives good results when applied to benchmarks mimicking the complexity of hierarchically nested complex systems. We call our algorithm statistically validated hierarchical clustering (SVHC)[2, 1]. Specifically, for each pair of parent and children nodes in the hierarchical tree, we test the difference between the proximity measure (in our approach a dissimilarity) associated with a clade h and the dissimilarity measure associated with the clade defined by its parent node in the genealogy of the dendrogram. The statistical test we perform consider as a null hypothesis that the dissimilarity of the parent node is larger than the dissimilarity of the children node. Our tests are performed by considering multiple hypothesis test correction. In fact, we always apply the control of false discovery rate. By selecting those clades that reject our null hypothesis, we identify a hierarchically nested partition involving a certain number of elements of the investigated systems. In order to evaluate the performance of our method, we test it with some benchmarks obtained by using a hierarchical factor model.

By performing numerical experiments on a representative benchmark and on a reference empirical dataset, we show that our algorithm is quite accurate and much faster and scalable than the state of the art algorithm (Pvclust). Moreover, it shed light on the role and limits of the presence or absence of a procedure for the multiple hypothesis test correction (Fig. 1). For these reasons, we believe the new algorithm will be of interest for those scholars working with large multivariate datasets in biology, computer science, neuroscience, physics, sociology, and other disciplines dealing with large scale multivariate data.



(a) (b) (c)

Fig. 1. (a) and (b) Numerical experiments with a benchmark composed by 12 overlapping clusters. (a) Number of statistically validated clusters detected by the algorithms as a function of the system size N. (b) Computational time of the algorithms as a function of the system size N; (c) hierarchical tree (average linkage HC) and correlation matrix of lung tissues dataset [3]. In the correlation matrix we highlight hierarchically nested clusters detected by our method.

Summary. We develop a greedy algorithm that is fast and scalable in the detection of a nested partition extracted from a dendrogram obtained from hierarchical clustering of a multivariate series. Our algorithm provides a p-value for each clade observed in the hierarchical tree. The p-value is obtained by computing a number of bootstrap replicas of the dissimilarity matrix and by performing a statistical test on each difference between the dissimilarity associated with a given clade and the dissimilarity of the clade of its parent node. By performing numerical experiments on a representative benchmark and on a reference empirical dataset, we show that our algorithm is quite accurate and much faster and scalable than the state of the art algorithm (Pvclust). Moreover, it shed light on the role and limits of the presence or absence of a procedure for the multiple hypothesis test.

References

1. SVHC code. https://github.com/cbongiorno/svhc
2. Bongiorno, C., Micciché, S., Mantegna, R.N.: Nested partitions from hierarchical clustering statistical validation. arXiv preprint arXiv:1906.06908 (2019)
3. Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., Van De Rijn, M., Rosen, G.D., Perou, C.M., Whyte, R.I., et al.: Diversity of gene expression in adenocarcinoma of the lung. Proceedings of the National Academy of Sciences 98(24), 13784–13789 (2001)
4. Suzuki, R., Shimodaira, H.: Pvclust: an r package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12), 1540–1542 (2006)

Embeddings-enhanced Language Communities Separation

Sandra Mitrović[1], Steven Skiena[2], and Jochen De Weerdt[1]

[1] LIRIS, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium,
sandra.mitrovic@kuleuven.be,
[2] Stony Brook University, Long Island, NY, USA

1 Introduction

The number of methods proposed for community detection (CD) in graphs is constantly increasing, however, those typically do not support setting the number of resulting communities upfront [1],[2]. Nevertheless, in real-life problems, the number of existing communities could be known and one "just" needs to assign instances to the most appropriate communities. One such problem that we are considering here is whether in a multi-language country it would be possible to separate language communities based on the (solely) customer calling data (hence without any additional attributes that might be more related to the language used). To this end, we propose using k-means clustering on top of learnt proximity preserving customer representations (and not hand-crafted topological features as has been a common practice so far).

Therefore, the main contribution of this study is providing the evidence that even the simplest clustering algorithms can perform better than some well-known, sophisticated CD methods, if they are applied on top of learnt representations. This is especially the case when only a network topology is available and no other seemingly related attributes could be derived from the underlying network.

2 Methodology and Experimental Setup

Methodology Our methodology consists of three main steps. The first one is a **call network construction**, implemented (classically) by assigning nodes to customers and adding links if corresponding customers had a call. We resort to exploiting the largest connected component (LCC) with ∼3.9M nodes and ∼5.7M edges as it achieves better performance than the original graph. Secondly, we perform **representation learning on the call network**, whereby we learn node (customer) representations (aka embeddings). To this end, we use a learning method based on random walks and word2vec [3], proximity preserving neural network language model. More concretely, we exploit both SkipGram and CBOW methods to generate embeddings. Finally, the third step is **clustering of the embeddings**. More specifically, we perform k-means clustering on previously generated embeddings, imposing the number of clusters to be equal to that of the ground truth. Finally, we evaluate the quality of clustering using the Adjusted Rand Index (ARI) and balanced accuracy (BACC), to account for the imbalanced number of instances in the ground truth clusters.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Data We operate with anonymized call detail records (CDRs) containing info about caller, callee, date/time and duration per call. In addition, for a subset of users we have their voicemail language settings. These serve as a proxy for the language spoken by a customer. Among four identified languages (denoted as Lang1-Lang4), Lang2 and Lang3 are far more used, so both 2 and 4 clusters were considered as a ground truth.

Baselines Three methods were used: 1) Louvain [1], a well-known CD algorithm originally proposed as well for distinguishing language communities, 2) the Asynchronous Fluid (*AF*) community detection algorithm [4], based on the interaction of fluids in an environment, that allows for a predefined number of communities (the main motivation for using it) and, 3) K-means clustering combined with various hand-crafted network features (denoted as *net_feat*) which would permit to directly compare the efficiency of learnt embeddings against the manually derived topology-based information.

3 Results

Applying Louvain [1] on the LCC, yielded as much as 2175 different communities. An analysis of the largest 50 of them (Figure 1), shows that typically one particular language (mostly Lang2 and Lang3) dominates each community. However, there is a problem of the same language users being scattered over many clusters. Furthermore,

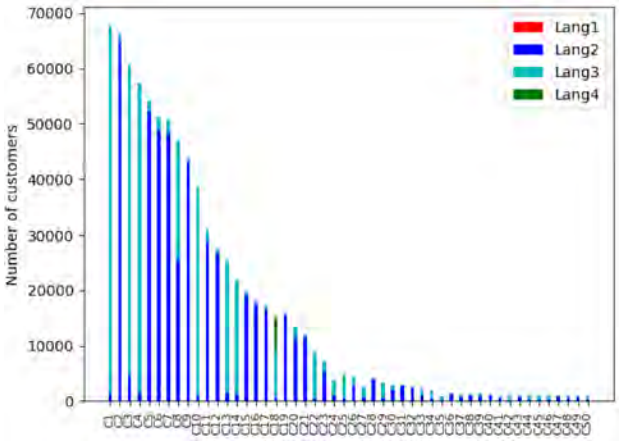


Fig. 1. Language distribution across the largest 50 Louvain communities.

Louvain does not allow specifying the number of communities, and merging them post-hoc does not guarantee (to say the least) maximal modularity (the Louvain's main idea).

The rest of the methods could be properly benchmarked given that the number of pre-set communities was the same (2 or 4). As can be seen from Table 1, both ARI and BACC are the highest when embeddings were used, with SkipGram providing better results than competing CBOW. Moreover, the ARIs close to 0.0 for methods based on

Table 1. ARI (the best in boldface) and BACC using 4 different methods and 2 versions of ground truth. Emb_{SG} and Emb_{CBOW} stand for embeddings obtained by SkipGram and CBOW methods, respectively, while net_feat refers to hand-crafted features derived from the LCC directly.

Method	ground truth: 4 clusters		ground truth: 2 clusters	
	ARI	BACC	ARI	BACC
Asynchronous Fluid	0.0413	0.2713	-0.0114	0.4735
k-means+net_feat	-0.0001	0.2500	-0.0001	0.4998
k-means+Emb_{SG}	**0.3651**	0.4075	**0.8002**	0.9466
k-means+Emb_{CBOW}	0.1736	0.3261	0.4297	0.8036

AF and *net_feat*, clearly indicate that the corresponding labelings are almost random. Furthermore, using only 2 clusters as ground truth provides better results (except for the *AF* method), probably as the two less used languages introduce some noise otherwise.

It is worth mentioning that due to the huge size of the LCC, calculating many features that deemed as potentially informative, was not feasible within a reasonable time frame (48 hours). This was particularly the case with most of the centrality measures. As such, the final set of features (per node) included in *net_feat* were: first-order degree, average neighbor degree, clustering coefficient, degree centrality, number of triangles and PageRank score. Similar computational issues were encountered with standard community detection algorithms such as [2].

Summary. The obtained results clearly demonstrate that embeddings can improve the quality of clustering and lead to outperforming sophisticated CD methods. This is especially valuable when there is a lack of any additional data (except for network topology, that is). Furthermore, presented method was proven to be scalable on large networks.

As a future work we envision benchmarking obtained results with the methods aiming at learning network (node) representations taking into account underlying communities such as [5]. Additionally, it would be worthwhile taking a further look into the dynamics related to observed communities.

References

1. Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, no. 10, P10008.
2. Aaron Clauset, Mark E.J. Newman and Cristopher Moore. 2004. Finding community structure in very large networks. Physical review E, 70(6), 066111.
3. Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. ArXiv preprint arXiv:1301.3781.
4. Ferran Parés, Dario Garcia Gasulla, Armand Vilalta, Jonatan Moreno, Eduard Ayguad, Jess Labarta, Ulises Corts and Toyotaro Suzumura. 2017. Fluid communities: a competitive, scalable and diverse community detection algorithm. In International Conference on Complex Networks and their Applications (pp. 229-240). Springer, Cham.
5. Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community preserving network embedding. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17) (pp. 203-209). AAAI Press.

Motif-Based Spectral Clustering of Weighted Directed Networks

William George Underwood[1,2] and Mihai Cucuringu[2,3]

[1] Princeton University, Princeton NJ 08544, USA
[2] University of Oxford, Oxford, UK
[3] The Alan Turing Institute, London, UK

1 Introduction

Networks are ubiquitous in modern society; from the Internet and online blogs to protein interactions and human migration, we are surrounded by inherently connected structures [4]. A fundamental problem in network analysis and machine learning is that of *clustering*, which aims to identify groups of nodes that are highly inter-connected or exhibit similar features. Spectral methods for network clustering have a long and successful history, and have become increasingly popular in recent years due to their computational efficiency and amenability to theoretical analysis under various probabilistic models. However, traditional spectral methods have shortcomings, which stem from their inability to capture latent higher-order network structures [2], and the challenges faced when handling directed edges [3], which renders the adjacency matrix no longer symmetric. Motif-based spectral methods have proven more effective for clustering directed networks on the basis of higher-order structures [7], with the introduction of the *motif adjacency matrix* (MAM). We explore motif-based spectral clustering methods with a focus on addressing these shortcomings for weighted directed networks, and augment our findings with numerical experiments on synthetic and real-world networks.

2 Problem Statement and Main Results

We consider clustering a weighted directed graph without self-loops or multiple edges. To exploit higher-order structures, we look for the occurrence of motifs (small connected subgraphs, Figure 1). We consider the weighted motif adjacency matrix M associated with a graph $\mathscr{G}$ and motif $\mathscr{M}$, where M_{ij} is the total weight of all instances of $\mathscr{M}$ in $\mathscr{G}$ containing both nodes i and j, and apply traditional spectral clustering to the resulting (symmetric) matrix M. For motifs on at most three nodes, Proposition 1 gives a fast and parallelizable matrix multiplication-based procedure for computing MAMs. In addition, we also present a novel motif-based method for clustering bipartite graphs.

Proposition 1 (MAM formula). *Suppose $\mathscr{G}$ is a graph on n vertices, and $\mathscr{M}$ is a motif on at most three vertices. Then calculating an MAM takes at most 18 matrix multiplications, 22 entry-wise multiplications and 21 additions of $n \times n$ matrices.*

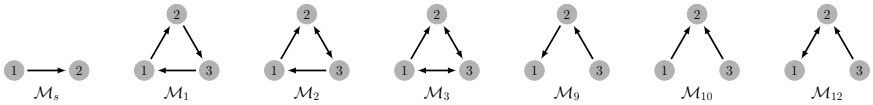


Fig. 1. Example of directed motifs which might appear as subgraphs of a larger graph.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2.1 Motif-Based Clustering in Directed Graphs

We consider a family of directed stochastic block models (DSBMs) that exhibit imbalanced flows in terms of the edges between clusters, and show that our motif-based method performs better than traditional spectral clustering. Figure 2 plots the popular Adjusted Rand Index (ARI) [6] attained by various motifs, averaged over 20 trials, for asymmetric two-block DSBMs with $n = 200$ nodes. The first motif $\mathscr{M}_s$ yields the traditional spectral clustering algorithm, while the others consider higher-order structures. The top of the plot shows $|C|$, the number of nodes clustered by the algorithm. Higher values of ARI and $|C|$ are better, and clearly motif $\mathscr{M}_1$ outperforms traditional methods.

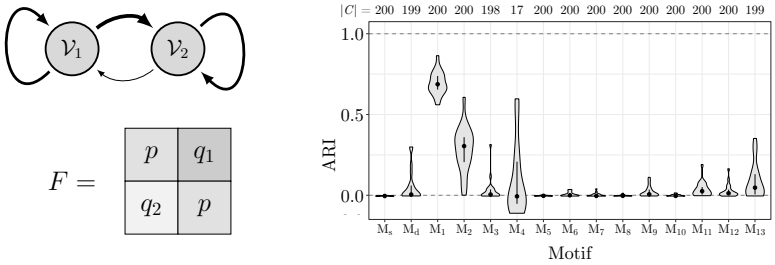


Fig. 2. Left: block structure and sparsity matrix of the asymmetric two-block DSBM. Right: ARI violin plot for the asymmetric two-block DSBM.

Next, we consider the US Migration data set [8], where the $n = 3107$ nodes denote the counties in mainland US, and the weighted directed edges indicate human migration flows during 1995–2000. Figure 3 shows the motif-based second eigenvector embeddings (x_2) and clusterings (C) obtained using various motifs, with $k = 7$ clusters.

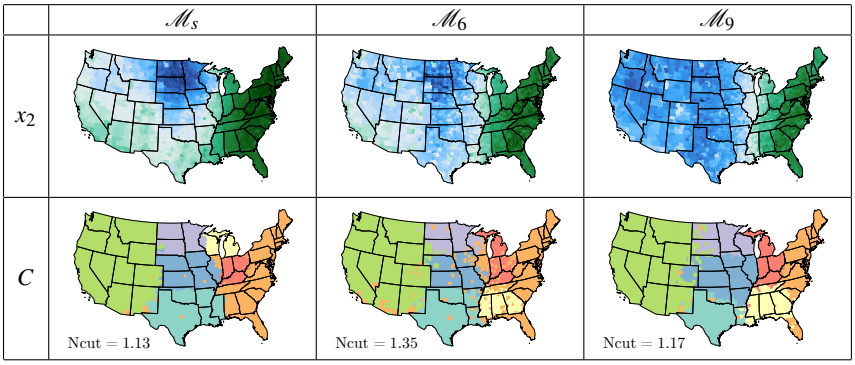


Fig. 3. Top: motif-based colorings of the US Migration network, from the second eigenvector of M. Bottom: clustering structure recovered from standard random-walk spectral clustering on M.

We also considered the US Political Blogs network [1], with $n = 1222$ nodes denoting blogs labelled as "liberal" or "conservative", and weighted directed edges indicating the number of citations between blogs. Figure 4 plots ARI against number of vertices clustered by various motifs, and shows the eigenvector embedding given by motif $\mathscr{M}_{12}$.

2.2 Motif-Based Clustering in Bipartite Directed Graphs

We consider bipartite stochastic block models (BSBMs), and show the effectiveness of motif-based methods for clustering them. We also demonstrate bipartite clustering on

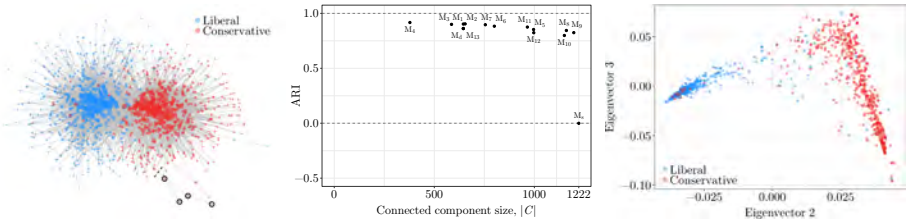


Fig. 4. Left: The US Political Blogs network. Middle: ARI versus largest connected component size for various motifs. Right: eigenvector embedding for motif $\mathscr{M}_{12}$, colored by political leaning.

the Unicode Languages network [5], where source nodes denote territories, destination nodes denote languages, and weighted directed edges indicate number of speakers. Figure 5 shows a clustering of territories into 6 clusters based on their common languages.

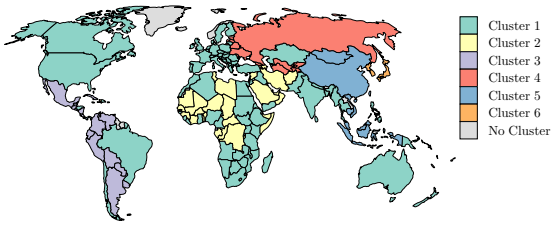


Fig. 5. Clustering of the territories from the Unicode Languages network.

3 Discussion and Conclusion

Motif-based spectral clustering is a valuable tool for clustering weighted directed networks, which is scalable and easy to implement. Potential extensions include an analysis of the differences between clustering based on functional and structural MAMs, a comparison with the Hermitian-based clustering in directed graphs [3], and application to directed core-periphery detection.

References

1. L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: Divided they blog. In *Proc. of the 3rd Intl. Workshop on Link Discovery*, pages 36–43. ACM, 2005.
2. A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.
3. M. Cucuringu, H. Li, H. Sun, and L. Zanetti. Hermitian matrices for clustering directed graphs: Insights and applications. *Submitted*, 2019.
4. E. D. Kolaczyk and G. Csárdi. *Statistical Analysis of Network Data with R*, volume 65. Springer, New York, 2014.
5. KONECT: The Koblenz Network Collection. Unicode Languages network dataset. http://konect.cc/networks/unicodelang. Accessed on 24 Mar 2019.
6. W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
7. C. E. Tsourakakis, J. Pachocki, and M. Mitzenmacher. Scalable motif-aware graph clustering. In *Proc. of the 26th Intl. Conference on World Wide Web*, pages 1451–1460, 2017.
8. U.S. Census Bureau. County-to-county migration flow files. https://www.census.gov/population/www/cen2000/ctytoctyflow/index.html, 2002. Accessed on 02 Mar 2019.

Finding meaningful communities in complex networks

Armin Pournaki, Felix Gaisbauer, Eckehard Olbrich, Sven Banisch

Max Planck Institute for Mathematics in the Sciences, Leipzig
pournaki@mis.mpg.de

Introduction

Community detection has grown to become a standard tool for the observation of modular structures [3]. One of the most used approaches is optimising the so-called modularity function [6], for which a wide range of algorithms exist [5, 2, 9]. Next to modularity based approaches, new graph-generating ideas based on stochastic block modeling emerged [8], allowing the discovery of disassortative community structures and even the inclusion of metadata in the partitioning [7].
Another way to detect communities can be considering coordination games on networks. The social feedback model [1] provides an agent-based interaction model in which agents change their opinion based on the reaction of their neighbors. In a binary opinion space, the stability of the final opinions (or communities) on the network can be assessed using the cohesion measure [4]. We will see that the per-node cohesion, which we call "node belongingness", is closely related to the modularity.
The following efforts aim to provide an interpretation of the modularity metric in relation to game-theoretic models. In this context, we want to ask how many opinions a certain network configuration supports, and what the roles of single nodes can be in enabling the emergence or disappearance of certain opinions in networks.

Social Feedback Model

There are different opinions o_i that agent i can adopt and express to their neighbors. Agents become more convinced of an opinion if the response from their neighbors is positive, and less convinced otherwise.
The internal evaluation of the agent is updated by:

$$\mathscr{C}_i(o) \leftarrow \begin{cases} (1-\alpha)\,\mathscr{C}_i(o) + \alpha r_i : & \text{if } o = \text{ expression} \\ \mathscr{C}_i(o) : & \text{else} \end{cases}$$

with reward $r_i = 1$ if $o_i = o_j$ and -1 else.
Agents express the opinion they most strongly support during the current timestep:

$$o_i = \arg\max_o \mathscr{C}_i(o)$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Modularity and mean belongingness

The modularity is defined as:

$$Q = \frac{1}{\sum_i k_i} \sum_{i,j} (A_{ij} - P_{ij})\ \delta\left(g_i g_j\right) \tag{1}$$

where k_i is the weight of node i, P_{ij} is a null model and g_i is the community index of node i. Usually, the null model is based on the degree distribution of the network, resulting in:

$$Q = \frac{1}{\sum_i k_i} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{\sum_i k_i})\ \delta\left(g_i g_j\right) \tag{2}$$

We define the belongingness of a node to its community as the fraction of its neighbors that share the same opinion:

$$c_i = \frac{\sum_j \delta(g_i g_j)\, A_{ij}}{\sum_j A_{ij}} \tag{3}$$

We compute the mean belongingness:

$$\frac{1}{N}\sum_i c_i = \frac{1}{N}\ \sum_i\ \frac{\sum_j \delta(g_i g_j)\, A_{ij}}{\sum_j A_{ij}} = \frac{1}{N}\ \sum_i\ \frac{\sum_j \delta(g_i g_j)\, A_{ij}}{k_i} \tag{4}$$

The modularity can be rewritten in the following way :

$$Q = \frac{1}{\sum_i k_i} \sum_{i,j} A_{ij}\ \delta\left(g_i g_j\right) - \kappa = \sum_i \frac{k_i}{\sum_j k_j} c_i - \kappa \tag{5}$$

The value κ depends on the null model P_{ij}:

$$\kappa = \frac{1}{\sum_i k_i}\ \sum_{i,j} \frac{k_i k_j}{2m}\ \delta(g_i g_j) = \frac{1}{\sum_i k_i}\ \sum_{i,j} P_{ij}\ \delta(g_i g_j) \tag{6}$$

We see in (5) that the modularity is proportional to the node belongingness. This is a first step in interpreting the modularity function in terms of game-theoretic stability criteria. For a space of two opinions, the stability criterion based on cohesion [4] is quite straightforward: an opinion configuration is stable if $\min_i(c_i) > 0.5$. However, if we consider more than two possible opinions, the criterion is not as clear. We aim to connect these notions of stability to a multi-opinion context, which could be one path to testing modularity-based partitions for meaning.

Results

We test the social feedback model on the paradigmatic Karate Club network [10]. Links in the network represent social ties between the members of a university karate club. The club is torn by a dispute between the instructor and the president, which eventually leads the network to split in two, resulting in the so-called "ground truth" partition. Except for member number 9, who picked his final faction stategically, Zachary was able

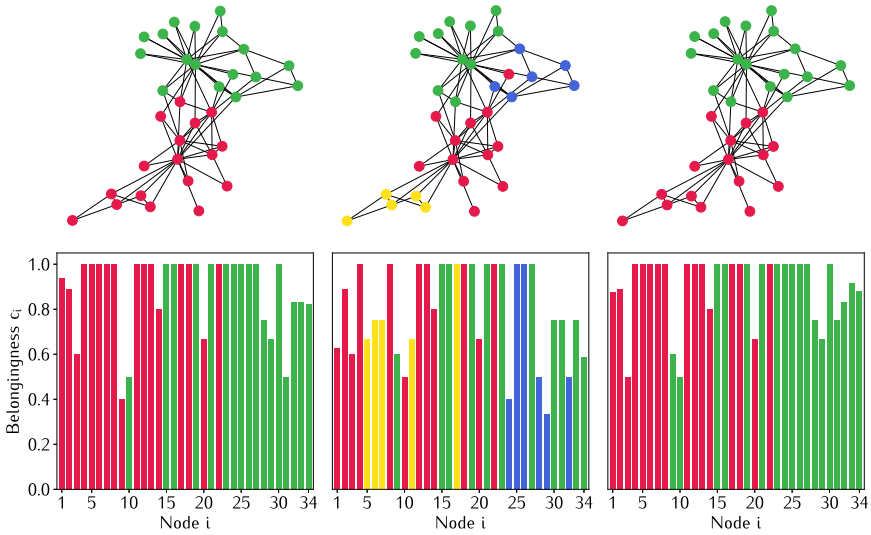


Fig. 1. Different partitions of the Karate Club network and their node belongingness distribution. Nodes are colored according to their community. The Karate Club network's "ground truth" partition is presented on the left. In the middle is a Louvain partition, which presents four communities. On the right is the outcome of the social feedback model, in which the individual opinions were initialised using the Louvain partition. Only two opinions survive. The learning rate is set to $\alpha = 0.015$. On the lower plots, each node's belongingness c is shown: $\min_{louvain} c_i = \frac{1}{3} < \min_{groundtruth} c_i = \frac{2}{5} < \min_{social\,feedback} c_i = \frac{1}{2}$.

to find the partition using a maximum flow algorithm [10].
We compute a Louvain partition of this network with resolution 1.0. The resulting partition is used to initialise the node's opinions for the social feedback model at $t = 0$. The learning rate is set to $\alpha = 0.015$. At $t = 100000$, only two opinions survive, corresponding to the ground truth partition except for member number 9. Figure 1 shows the different partitions of the network and the according node belongingness values for each node. The Louvain partition presents several nodes with $c_i < 0.5$. From this first look, it is possible to get an intuition of stability of certain partitions. For instance, the absorbtion of two out of four communities from the Louvain partition could be predicted using the node belongingness value.

Summary and Outlook

This abstract provides a first approach at interpreting the abstract modularity function in the game-theoretic context of the social feedback model. The model can be used to test a given network partition for stability, which is demonstrated on the Karate Club network. Here, two small commmunities are absorbed by larger ones, resulting in a two-opinion network.

Further work will investigate the role of individual nodes in an opinion space, especially at the borders of communities, to see if there are interaction patterns that enhance or suppress opinion absorbtion. Ultimately, the question of how many communities are supported given a network structure will be addressed in the course of this research.

References

1. Banisch, S., Olbrich, E.: Opinion polarization by learning from social feedback. The Journal of Mathematical Sociology 43(2), 76–103 (2018), https://doi.org/10.1080/0022250x.2018.1517761
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008), http://stacks.iop.org/1742-5468/2008/i=10/a=P10008
3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75–174 (Feb 2010), https://doi.org/10.1016/j.physrep.2009.11.002
4. Morris, S.: Contagion. The Review of Economic Studies 67(1), 57–78 (2000)
5. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical Review E 69(6) (2004), https://doi.org/10.1103/physreve.69.066133
6. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006), https://doi.org/10.1073/pnas.0601602103
7. Newman, M.E.J., Clauset, A.: Structure and inference in annotated networks. Nature Communications 7(1), 11863 (2016), https://doi.org/10.1038/ncomms11863
8. Peixoto, T.: Bayesian stochastic blockmodeling (5 2017), 42 pages, 16 figures, Chapter in "Advances in Network Clustering and Blockmodeling", edited by P. Doreian, V. Batagelj, A. Ferligoj, (Wiley, New York, 2018 [forthcoming])
9. Traag, V.A., Waltman, L., van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Scientific reports 9 (2019)
10. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33(4), 452–473 (Dec 1977), https://doi.org/10.1086/jar.33.4.3629752

Handling Noisy Constraints in Semi-supervised Overlapping Community Finding

Elham Alghamdi, Ellen Rushe, Mehran H.Z. Bazargani,
Brian Mac Namee, and Derek Greene

School of Computer Science, University College Dublin, Ireland

1 Introduction

Community structure is an essential property that helps us to understand the nature of complex networks. Since algorithms for detecting communities are unsupervised in nature, they can fail to uncover useful groupings, particularly when the underlying communities in a network are highly overlapping [1]. Recent work has sought to address this via semi-supervised learning [2], using a human annotator or "oracle" to provide limited supervision. This knowledge is typically encoded in the form of must-link and cannot-link constraints, which indicate that a pair of nodes should always be or should never be assigned to the same community. In this way, we can uncover communities which are otherwise difficult to identify via unsupervised techniques.

However, in real semi-supervised learning applications, human supervision may be unreliable or "noisy", relying on subjective decision making [3]. Annotators can disagree with one another, they might only have limited knowledge of a domain, or they might simply complete a labeling task incorrectly due to the burden of annotation. Thus, we might reasonably expect that the pairwise constraints used in a real semi-supervised community detection task could be imperfect or conflicting. The aim of this study is to explore the effect of noisy, incorrectly-labeled constraints on the performance of semi-supervised community finding algorithms for overlapping networks. Furthermore, we propose an approach to mitigate such cases in real-world network analysis tasks. We treat noisy pairwise constraints as anomalies, and use an autoencoder, a commonly-used method in the domain of anomaly detection, to identify such constraints. Initial experiments on synthetic network demonstrate the usefulness of this approach.

2 Methods and Experimental Design

The key aspect of our work is an iterative approach using an autoencoder to remove noisy pairwise constraints selected by the AC-SLPA algorithm [2]. An *autoencoder* (AE) refers to a neural network architecture that attempts to reconstruct a given input in an effort to learn an informative latent feature representation. Formally, for an input vector x, we attempt to map x to a reconstruction of itself x'. By doing this, a latent representation of the data is created in the hidden layer(s) of the network [4]. These networks can utilize a "bottleneck" configuration where the hidden layer(s) of the network compress the data [4]. The network is trained by minimizing the mean squared error (MSE) between the reconstruction and input. Additionally, autoencoders can be constrained to

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

enforce sparsity in the network and therefore no longer require a compressed network capacity. One type of constrained autoencoder adds a sparsity penalty to hidden representations by constraining their absolute value. This penalty term is weighted and added to the cost function. In our work we employ the above neural network architecture to identify potentially noisy pairwise constraints selected by AC-SLPA before applying the community detection process.

Firstly, feature vectors are constructed as inputs to the autoencoder, one vector per input constraint pair. Along with the constraint type, the other features include standard measures based directly on the network topology: whether the pair of nodes shares an edge, their number of common neighbors, shortest path length, and cosine similarity. We also include more complex features: their SimRank similarity [6] and their similarity as computed on a *node2vec* embedding generated on the network [5]. From this data, the model then learns to reconstruct the original constraints from the latent representation. The reconstruction error is then given by the difference between the original constraints and the reconstruction. A large error is indicative of an anomaly (i.e. a noisy constraint), while a low error indicates a "normal" example (i.e. a correctly-labelled constraint). The expectation is that, as the vast majority of pairwise constraints are non-noisy, the autoencoder's latent representation will be biased towards these examples. This makes the model somewhat robust to outliers. Based on this property, it is then assumed that examples which are noisy will have a high reconstruction error.

As our initial evaluation, we assess the capability of autoencoders to detect noisy constraints. Once the set of constraints is selected by AC-SLPA and labeled by the oracle, the autoencoder is trained on this set. These are then passed through the autoencoder once again to obtain a reconstruction error for each constraint. The AUC over this error is calculated, which provides an estimate of the number of constraints that were successfully detected in the absence of a definitive threshold. The number of layers in each autoencoder is varied to examine whether this task benefits from a deeper model. We consider both compression-based autoencoders and sparse autoencoders.

Evaluations are performed on 64 LFR benchmark networks containing either small or large communities, for a variety of parameters $\{N, Om, On, \mu\}$ (see Table 1). The depth of the autoencoder is varied to assess its effect on performance. In the case of the compression autoencoders, the nodes are gradually decreased in the encoder and increased in the decoder, while this compression is not necessary for the L1 constrained models [4]. In the case of the constrained autoencoders, the sparsity weight is kept at 10^{-5}. All models were trained with a learning rate of 10^{-3} for a maximum of 100 epochs and a batch size of 256.

3 Results

The results in Table 1 are divided into two parts, which represent the AUC scores of the autoencoder on networks with 10% and 50% overlapping nodes respectively, averaged across 10 runs. Each table entry shows the AUC value of an AE model (on the rows) for each network (on the columns). For each network, the AUC scores of AE models are ranked, and the best performance is highlighted in bold. The last column reports the average rank score of each model. As we can see, all AE models show high AUC scores,

Table 1: AUC scores on LFR networks with 10% of noise in pairwise constraints. AE* [layers dimension]: indicates the number of layers in compression autoencoders, and AE*_l1 [layers dimension]: indicates the number of layers in L1 constrained autoencoders: AE1: [7,3,7], AE1_L1: [7,7,7], AE2: [7,5,3,5,7], AE2_L1: [7,7,7,7,7], AE3: [7,6,5,3,5,6,7], AE3_L1: [7,7,7,7,7,7,7].

(a) AUC scores on networks with 10% overlapping nodes

Comm. size	Large Communities								Small Communities								Average Rank
μ	0.1				0.3				0.1				0.3				
Om	2	4	6	8	2	4	6	8	2	4	6	8	2	4	6	8	
AE1	0.752	0.736	0.777	0.736	0.829	0.770	0.744	0.751	0.775	0.757	0.756	0.773	0.795	0.733	0.773	0.739	4.4 (4)
AE1_L1	0.759	0.800	0.801	0.758	**0.837**	0.772	**0.829**	0.732	**0.832**	0.828	0.774	0.766	**0.810**	0.824	0.826	0.759	2.9 (2)
AE2	0.760	0.739	**0.803**	0.776	0.797	0.787	0.798	0.749	0.783	0.795	0.765	0.764	0.786	0.780	0.775	0.773	3.3 (3)
AE2_L1	**0.762**	0.706	0.791	0.760	0.792	**0.801**	0.789	0.784	0.775	0.795	0.798	0.769	0.770	0.834	**0.831**	0.813	2.9 (2)
AE3	0.754	**0.809**	0.771	**0.810**	0.794	0.797	0.796	**0.792**	0.817	**0.833**	**0.836**	**0.822**	0.769	**0.839**	0.827	**0.849**	**2.3 (1)**
AE3_L1	0.720	0.777	0.773	0.776	0.779	0.751	0.764	0.740	0.729	0.753	0.726	0.793	0.786	0.795	0.774	0.782	4.4 (4)

(b) AUC scores on networks with 50% overlapping nodes

Comm. size	Large Communities								Small Communities								Average Rank
μ	0.1				0.3				0.1				0.3				
Om	2	4	6	8	2	4	6	8	2	4	6	8	2	4	6	8	
AE1	0.744	0.797	0.823	0.793	**0.832**	0.815	0.804	**0.798**	0.779	**0.829**	0.828	0.778	0.813	0.831	0.849	0.806	3.3 (3)
AE1_L1	0.766	0.804	0.811	**0.834**	0.788	0.788	0.826	0.766	**0.847**	0.755	**0.874**	**0.818**	0.756	0.826	0.842	**0.836**	3.1 (2)
AE2	0.741	0.767	0.799	0.668	0.771	0.806	0.803	0.676	0.783	0.760	0.784	0.799	0.762	0.808	0.823	0.812	4.9 (5)
AE2_L1	**0.780**	0.798	0.791	0.833	0.794	**0.827**	0.818	0.783	0.801	0.776	0.867	0.790	0.792	**0.879**	0.848	0.819	**2.6 (1)**
AE3	0.696	0.720	0.706	0.757	0.782	0.745	0.823	0.779	0.822	0.770	0.858	0.669	**0.820**	0.837	0.808	0.803	4.4 (4)
AE3_L1	0.752	**0.824**	**0.835**	0.805	0.831	0.808	**0.837**	0.774	0.787	0.785	0.860	0.669	0.801	0.853	**0.852**	0.811	**2.6 (1)**

with the lowest scores around 70%. However, we see the AE3 models perform better on networks with $On = 10\%$, while AE1_L1 and AE2_L1 also perform well here. On the networks with $On = 50\%$, AE2_L1 and AE3_L1 are the top-ranked models. In general, these results suggest that deeper autoencoder models do not perform significantly better than simpler ones when detecting noisy constraints.

In summary, our proposed approach currently yields promising results on benchmark networks. A second set of experiments is currently in progress, which directly evaluates the performance of AC-SLPA when incorporating reliable constraints as selected by the autoencoder model, on both synthetic and real-world networks.

References

1. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature 466(7307), 761–764 (2010)
2. Alghamdi, E., Greene, D.: Active semi-supervised overlapping community finding with pairwise constraints. Applied Network Science 4 (2019)
3. Amini, M.R., Gallinari, P.: Semi-supervised learning with an imperfect supervisor. Knowledge and Information Systems 8(4), 385–413 (2005)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
5. Grover, A., Leskovec, J.: Node2vec: Scalable feature learning for networks. In: Proc. SIGKDD'16. pp. 855–864. ACM (2016)
6. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proc. SIGKDD'02. pp. 538–543. ACM (2002)

Community Detection in Interval-Weighted Networks

Hélder Alves[1], Paula Brito[2], and Pedro Campos[2]

[1] FCUP & LIAAD INESC TEC, University of Porto, Portugal
[2] FEP & LIAAD INESC TEC, University of Porto, Portugal

1 Introduction

In the classical network theory, in weighted (or valued) networks, edge weights are constants [1, 2]. However, in real–world applications these weights may vary within ranges rather than being constants [3]. To better model such variability of weights in a network, instead of using constants or point values (real numbers) and associated methods to represent the information present in the edges of the networks, we represent weights as intervals [4]. A representation of these values in the form of closed intervals composed with the precise information following the *ontic* approach [5], can be more meaningful and useful in a dynamic environment than point–valued output, as these intervals contain more information in expressing raw data variability [6].

Although several extensions of modularity to weighted networks were proposed, none takes into account the variability of link weights. To fill this gap, we extend both, Newman's modularity for weighted networks [2], and one state–of–the–art greedy method to optimize modularity introduced by Blondel *et. al.* [7] (the Louvain algorithm), to the general case of *interval–weighted networks* (IWN). Finally, we apply our community detection approach for IWN to a real–world commuter network between the Portuguese mainland municipalities.

2 Modularity in Interval-Weighted Networks

The generalization of the (unstandardised) modularity for weighted networks, $Q^W = \sum_{C\in\mathscr{C}}\sum_{i,j\in C}(o_{ij}-e_{ij})$ (where $\mathscr{C}$ is a partition of the vertices into q sets), gain of modularity ($\Delta Q^W = Q^W_{new} - Q^W_{last}$) and consequently of the Louvain method to this new approach, was done considering that the IWN can be represented as a *contingency table*, denoted by O^I, whose cells represent the *observed interval–weights* $o^I_{ij} = [\underline{o}_{ij}, \overline{o}_{ij}]$ ($\overline{o}_{ij} \geq \underline{o}_{ij} > 0$; $o^I_{ij} \subseteq \mathbb{R}^+$), if there is an weighted edge between vertices (i, j) and zero otherwise. The *interval total weight/strength* attached to vertex i, is denoted by $s^{IO}_i = \sum^n_{j=1}[\underline{o}_{ij}, \overline{o}_{ij}]$, and the *total weight* is, $\sum^n_{i=1} s^{IO}_i = \sum^n_{j=1} s^{IO}_j = \sum^n_{i=1}\sum^n_{j=1}[\underline{o}_{ij}, \overline{o}_{ij}]$ (to simplify, hereafter we will use the notation $[2\underline{w}, 2\overline{w}]$). Analogously, and assuming independence between the vertices, the contingency table for the *expected interval–weights* is defined as $E^I = e^I_{ij}$, where e^I_{ij} is the interval–weight that would be obtained if the hypothesis of row–column independence were true, $e^I_{ij} = \left[\frac{\underline{s}^{IO}_i \underline{s}^{IO}_j}{2\overline{w}}, \frac{\overline{s}^{IO}_i \overline{s}^{IO}_j}{2\underline{w}}\right]$, $(0 \notin [2\underline{w}, 2\overline{w}])$. Further, these expected frequencies must pass through an *"adjustment"* of its total lower $(2\underline{w})$ and upper limits $(2\overline{w})$. The generalization of modularity (Q^W) and modularity gain

(ΔQ^W) to *interval data* was done as follows: assuming that we have a fixed partition consisting in two communities C_r and C_s, the *modularity for interval–weighted networks* is equal to: $Q^{IW} = \sum_r D(o_{rr}, e_{rr})$, where "D" represents the difference between the observed o_{rr} and the expected e_{rr} interval–weights; likewise, to evaluate the modularity gain resulting from the merging of the two communities C_r and C_s into a single community $C_t = C_r \cup C_s$, the *modularity gain for interval–weighted networks* is equal to: $\Delta Q^{IW} = Q^{IW}_{new} - Q^{IW}_{last}$. Then following the same procedure, Newman's normalization of modularity [9] was generalized for the case of IWN by: $Q^{IW}_{norm} = \frac{Q^{IW}}{Q^{IW}_{\max}} = \frac{\sum_r D(o_{rr}, e_{rr})}{D\left([2\underline{w}, 2\overline{w}], \sum_r e_{rr}\right)}$.

In the previous generalizations we face two major setbacks: *interval dependency*; and the fact that *the value of the distance between intervals is always positive*. To contour these drawbacks we propose the following three types of measures to evaluate de difference between two intervals $[\underline{x}, \overline{x}]$ and $[\underline{y}, \overline{y}]$: $d_1\left([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]\right) = \max\left\{\underline{x} - \underline{y}, \overline{x} - \overline{y}\right\}$, $d_2\left([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]\right) = \max\left\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\right\}$ *sign argmax*$\left\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\right\}$, and a "vectorial difference" $\vec{d}_3\left([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]\right) = \left(\underline{x} - \underline{y}, \overline{x} - \overline{y}\right)$. According to the type of difference used, other alternative modularity measures were defined. Similarly, various community detection methods based on the Louvain algorithm have also been developed.

3 Application to a Commuters Interval-Weighted Network

We analyse the community structure that emerges from the movements of daily commuters in mainland Portugal between the twenty three Regions NUTS 3 (Nomenclature of Territorial Units for Statistics) [8]. The applied methodology is capable of detecting productive regions composed of cohesive NUTS 3 in terms of commuting flows. The elements o^I_{ij} denote the maximum variability of the *bi–directional* flows ij and ji between the NUTS i and j (Figure1b): $o^I_{ij} = \left[\min\{\underline{o}'_{ij}, \underline{o}''_{ji}\}, \max\{\overline{o}'_{ij}, \overline{o}''_{ji}\}\right] = \left[\underline{o}_{ij}, \overline{o}_{ij}\right]$ (flows greater than 50 daily movements). Therefore, taking into account the assumption of *regular bi–directional* movements along the edges, the adjacency matrix is symmetric, $o^I_{ij} = o^I_{ji}$, and the network is described as an *undirected interval–weighted network*.

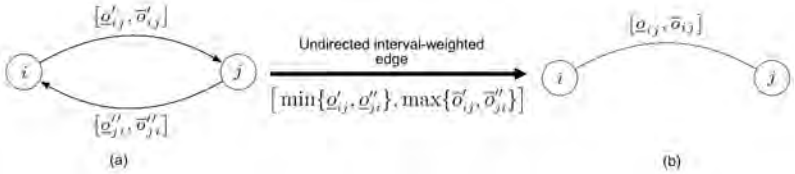


Fig. 1. (a) Bidirectional interval flows $i \to j$ and $j \to i$, (b) Undirected interval flow between ij.

For the sake of simplicity, we only report the results for the *difference* d_2. The final clustering reveals the existence of three NUTS 3 communities, with normalised modularity $Q^{IW}_{norm} = 0.596$ ($Q^{IW}_{\max} = 10792.1$, and $Q^{IW} = 6371.6$), which means a moderate/strong clustering structure. The Louvain algorithm for IWN reached maximum modularity at the end of the second pass. These communities roughly represent the division of the country into two major regions, the northern region (C_2: AMI, ATA, AMP, AVE, CAV, DOU, RAV, RCO, TES, TTM, VDL) and the southern region (C_1: ACE, AAL, BAL, ALI, ALG, AML, LTJ, OES, MTJ, RLE). However, the less "important" region, centre interior

of Portugal (C_3: BBA, BSE), forms a community of its own. Table 1 and Figure 2 below, show the *adjacency matrix* for the interval–weighted network and geographical representation that outcomes from this community detection method for IWN. These intervals account for the maximum variation in daily commuters flows within and between those communities.

Table 1. Interval–weighted adjacency matrix[a].

	C_1 ACE, AAL, BAL, ALI, ALG, AML, LTJ, OES, MTJ, RLE	**C_2** AMI, ATA, AMP, AVE, CAV, DOU, RAV, RCO, TES, TTM, VDL	**C_3** BBA, BSE
ACE, AAL, BAL, ALI, ALG, AML, LTJ, OES, MTJ, RLE	$[2562, 24720]$	$[966, 3483]$	$[269, 411]$
AMI, ATA, AMP, AVE, CAV, DOU, RAV, RCO, TES, TTM, VDL	$[966, 3483]$	$[4328, 41994]$	$[221, 731]$
BBA, BSE	$[269, 411]$	$[221, 731]$	$[110, 996]$

[a] NUTS 3: ACE-Alentejo Central, ALI-Alentejo Litoral, ALG-Algarve, AAL-Alto Alentejo, AMI-Alto Minho, ATA-Alto Tâmega, AML-Área Metropolitana de Lisboa, AMP-Área Metropolitana do Porto, AVE-Ave, BAL-Baixo Alentejo, BBA-Beira Baixa, BSE-Beiras e Serra da Estrela, CAV-Cávado, DOU-Douro, LTJ-Lezíria do Tejo, MTJ-Médio Tejo, OES-Oeste, RAV-Região de Aveiro, RCO-Região de Coimbra, RLE-Região de Leiria, TES-Tâmega e Sousa, TTM-Terras de Trás-os-Montes, VDL-Viseu Dão Lafões.

Fig. 2. Geographical representation.

Summary. We consider *Interval–Weighted Networks* (IWN) where the weights are represented by closed intervals, thus taking into account the variability of network edge weights. Accordingly, both Newman's modularity (Q), and modularity gain (ΔQ) for weighted networks, as well as Louvain's algorithm, were generalized to the general case of IWN. Further measures have been developed to evaluate the difference between the observed and expected values. Finally, we apply our community detection approach for IWN to a real–world commuter network between the Portuguese mainland municipalities to put in evidence homogeneous groups (communities) of territorial units.

References

1. Wasserman, S. and Faust, K.: Social Network Analysis: Methods and Applications, Cambridge University Press (1994).
2. Newman, M. E. J.: Analysis of weighted networks. Phys. Rev. E 70(5), 113 (2004).
3. Hu, C., Hu, P.: Interval-Weighted Graphs and Flow Networks. Knowledge Processing with Interval and Soft Computing, pp. 167-182, Springer-Verlag, London Limited (2008).
4. Moore, R.E., Kearfott, R.B, and Cloud, M.J.: Introduction to Interval Analysis, SIAM (2009).
5. Couso, I., Dubois, D.: Statistical reasoning with set-valued information: Ontic vs. epistemic views. International Journal of Approximate Reasoning 55(7) 15011518 (2014).
6. Hu, C., Kearfott, R.B., de Korvin, A. and Kreinovich, V.: Knowledge Processing with Interval and Soft Computing. Springer Science & Business Media (2009).
7. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Etienne, L.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008).
8. INE: Censos Resultados definitivos. Portugal - 2011 (2012).
9. Newman, M.E.J.: Networks: an introduction. New York: Oxford University Press (2010).

Acknowledgments: "This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project : UID/EEA/50014/2019".

Latent geometry inspired graph dissimilarities can boost community detection in complex networks

Alessandro Muscoloni[1], Claudio Durán[1] and Carlo Vittorio Cannistraci[1,2]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy

Abstract

Latent geometry has been recently shown to be relevant in applied fields of network science such as community detection and greedy routing [1]. However, there have been no general investigations so far on the extent to which latent geometry inspired graph dissimilarities can boost the task of community detection regardless of the particular type of principle adopted in the graph partitioning algorithm (stochastic flow, message passing, modularity optimization, etc...). For instance, Affinity propagation (AP) [2] and Markov Clustering (MCL) [3] are among the most effective algorithms for data clustering in high-dimensional feature space. However the numerous attempts to test their performance for community detection in real complex networks have been attaining results very far from the state of the art methods such as Infomap [4] and Louvain [5]. Indeed, the crucial problem is to convert the network topology in a 'smart-enough' pre-weighted connectivity or dissimilarity matrix that is able to properly address the algorithmic procedure behind these clustering techniques. Here we discuss how to leverage network latent geometry notions in order to design weighted matrices for community detection. Our results demonstrate that the dissimilarity measures we designed can boost AP [6], MCL and also Louvain, not only on several original real networks, but also when their structure is corrupted by noise artificially induced by missing or spurious connectivity. On the other side, further investigations are needed for enhancing Infomap. Finally, the results obtained on real networks are also con-firmed in tests performed on synthetic networks generated according to a hyperbolic latent geometry model [7] that induces community structure.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2

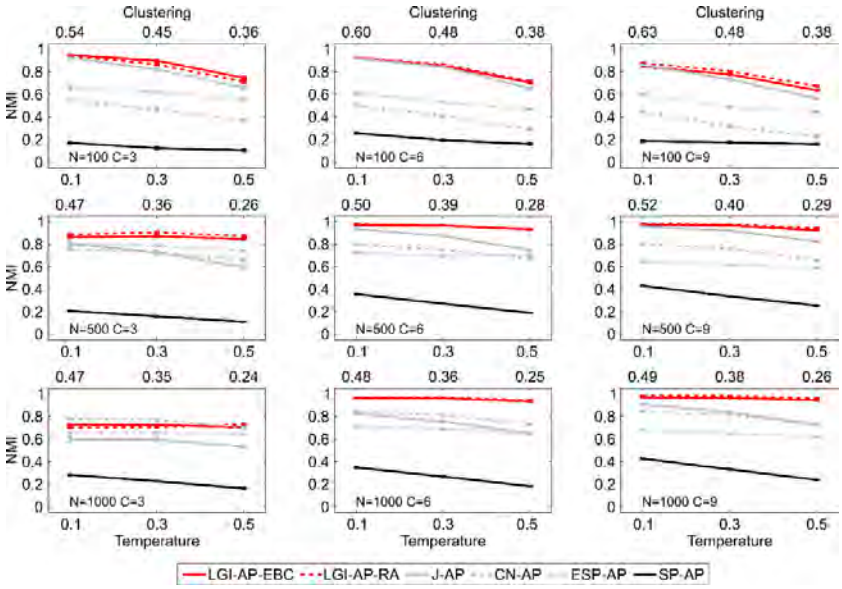


Fig. 1. Community detection on nPSO networks: comparison between different affinity propagation variants. Synthetic networks have been generated using the nPSO model with parameters $\gamma = 3$ (power-law degree distribution exponent), $m = 7$ (half of average degree), $T = [0.1, 0.3, 0.5]$ (temperature, inversely related to the clustering coefficient, whose respective value is reported on the upper part of each plot), $N = [100, 500, 1000]$ (network size) and $C = [3, 6, 9]$ (communities). For each combination of parameters, 100 networks have been generated. For each network the community detection methods LGI-AP-RA, LGI-AP-EBC, J-AP, CN-AP, ESP-AP and SP-AP have been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots report for each parameter combination the mean NMI and standard error over the random repetitions. For further details, please see the Reference [6].

Table 1. The table reports the Normalized Mutual Information (NMI) computed between the ground truth communities and the ones detected by every community detection algorithm for 8 real networks. NMI = 1 indicates a perfect match between the two partitions of the nodes. For Affinity propagation (AP) different dissimilarity matrices are compared: the best latent geometry inspired (LGI) variant and the ones introduced in previous studies, i.e. Jaccard (J), Common Neighbours (CN), Shortest Path (SP) or Euclidean Shortest Path (ESP). For further details, please see the Reference [6]. For Markov Clustering (MCL) the best latent geometry inspired (LGI) variant is compared with the unweighted version. The respective variants for each of the two methods are ranked by mean performance over the dataset.

Method	Karate	Opsahl 8	Opsahl 9	Opsahl 10	Opsahl 11	Polbooks	Football	Polblogs	mean NMI
LGI-AP	0.67	0.52	0.42	1.00	0.93	0.56	0.91	0.69	**0.71**
J-AP	0.73	0.48	0.45	1.00	0.96	0.39	0.89	0.40	0.66
ESP-AP	0.57	0.38	0.35	0.96	0.96	0.50	0.92	0.47	0.64
CN-AP	0.16	0.40	0.54	0.89	0.72	0.52	0.91	0.68	0.60
SP-AP	0.83	0.50	0.20	0.65	0.09	0.46	0.63	0.29	0.46
LGI-MCL	0.83	0.59	0.39	1.00	0.96	0.57	0.93	0.00	**0.66**
MCL	0.73	0.55	0.43	1.00	0.68	0.57	0.93	0.00	0.61

References

1. A. Muscoloni, J. M. Thomas, S. Ciucci, G. Bianconi, and C. V. Cannistraci, “Machine learning meets complex networks via coalescent embedding in the hyperbolic space,” Nat. Commun., vol. 8, 2017.
2. B. J. Frey and D. Dueck, “Clustering by Passing Messages Between Data Points,” Science, vol. 315, no. 5814, pp. 972–976, 2007.
3. S. van Dongen, “Graph clustering by flow simulation,” Graph Stimul. by flow Clust., 2000.
4. M. Rosvall and C. T. Bergstrom, “Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems,” PLoS One, vol. 6, no. 4, p. e18209, 2011.
5. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech Theory Exp., vol. 2008, no. 10, p. 10008, 2008.
6. A. Muscoloni and C. V. Cannistraci, “Latent Geometry Inspired Graph Dissimilarities Enhance Affinity Propagation Community Detection in Complex Networks,” arXiv:1804.04566, 2018.
7. A. Muscoloni and C. V. Cannistraci, “A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities,” New J. Phys. 20 052002, 2018.

Evaluation of pervasive community detection

Hiroshi Okamoto[1] and Xu-le Qiu[2]

[1] Department of Bioengineering, The University of Tokyo. Tokyo 113-8656 JAPAN,
okamoto@coi.t.u-tokyo.ac.jp
[2] Research & Development Group, Fuji Xerox Co., Ltd. Kanagawa 220-8668 JAPAN

1 Introduction

Exploiting random walk is an effective approach to designing methods for detecting communities from networks [1–3]. Following this approach, we have recently proposed a probabilistic machine-learning formulation of community detection, which we have called modular decomposition of Markov chain (MDMC) [4, 5]. This formulation postulates decomposition of an infinite random walk spreading over the entire network, from which we wish to detect communities, into local modules as proxy for communities. The decomposition is mathematically expressed by the mixture distribution:

$$p(n) = \sum_{k=1}^{K} \pi(k) p(n|k) \ , \tag{1}$$

where N is the total number of nodes; $p(n)$ is the 'global' probability that a random walker is at node n [6]; $p(n|k)$ is the 'local' probability that he/she is at node n conditioned that he/she is staying in community k; $\pi(k)$ is the probability that he/she is staying in community k. We have derived the EM algorithm to infer $p(n|k)$ and $\pi(k)$ [4, 5], by which community detection is attained.

The structure of each community k detected by MDMC is delineated by $p(n|k)$, which defines the relative strength of membership of each node n in community k. Since $p(n|k)$ normally takes a non-negative graded value, such a community has no clear boundary that separates members and non-members of the community. Such a structure of communities is described as "pervasive" [7]. Thus, MDMC detects communities as pervasively structured objects.

The present study is devoted to demonstrating that pervasive community detection, which is out of reach of most existing methods [7], is a key advantage of MDMC. First, we propose to use a specific type of stochastic block modelling to synthesize benchmark networks planted with pervasive communities. Then, MDMC's performance of pervasive community detection is quantitatively evaluated using these benchmark networks.

2 Methods

Benchmark networks planted with pervasively structured communities are mathematically synthesized using Ball-Karrer-Newman's stochastic block model (BKN's SBM) [8], which defines the probability of generating a network with adjacency matrix $\mathbf{A} =$

(A_{nm}) by Poisson distribution in the form

$$p(\mathbf{A}) = \prod_{n,\, m=1}^{N} \left[\frac{\left(\sum_{k=1}^{K_*} \theta_{nk}\theta_{mk}\right)^{A_{nm}}}{A_{nm}!} \exp\left(-\sum_{k=1}^{K_*} \theta_{nk}\theta_{mk}\right) \right] . \tag{2}$$

Here, K_* is the number of planted communities; θ_{nk} is a parameter representing the "propensity" of node n to block k and takes a continuous non-negative value, whereby delineating the pervasive structure of block k (namely, planted community k); $\sum_{k=1}^{K_*} \theta_{nk}\theta_{mk}$ is the rate for a Poisson event of generating a link between nodes n and m.

The $\{\theta_{nk}\}_{n=1}^{K_*}$ is stochastically generated so that they follow a power-law distribution $p(\theta_{nk}) \sim \theta_{nk}^{-\gamma}$. Assuming the power law stems from the observation that the degree distribution obeys a power law in many of real-world networks. The parameter values for the synthesis are set as follows: $N = 1000$; $K_* = 10$; $\gamma = 3$.

To quantitatively measure how correctly planted pervasive communities are recovered by community detection, we introduce the "maximum rank correlation (MaxRC)", defined as follows. Let $R_*(k)$ and $R(k')$ be the rank order of nodes defined in planted community k and detected community k' according to the descending order of θ_{nk} and $p(n|k')$, respectively. Spearman's rank correlation between $R_*(k)$ and $R(k')$, denoted by $r(R_*(k), R(k'))$, is then calculated for all combinations of k $(k = 1, \cdots, K_*)$ and k' $(k' = 1, \cdots, K)$. Therefore, MaxRC is given by

$$\mathrm{MaxRC} = \frac{1}{K^*} \sum_{k=1}^{K_*} \max_{k'} r\left(R_*(k), R(k')\right) . \tag{3}$$

BKN's SBM can also be used to detect pervasive communities. This is achieved by inferring θ_{nk} for the adjacency matrix $\mathbf{A} = (A_{nm})$ of a given network [8]. Indeed, BKN's SBM is one of few existing methods that can detect pervasive communities. Therefore, BKN's SBM is taken as a baseline for quantitative evaluation of MDMC's performance of pervasive community detection.

3 Results and Discussion

Each planted community k is delineated by θ_{nk}, or equivalently, 'normalized' propensity defined by $p_*(n|k) \equiv \theta_{nk} / \sum_{n=1}^{N} \theta_{nk}$. Panels in Fig. 1a show the normalized propensities for planted communities of the same network but with the node number (#) n sorted in descending order of $p_*(n|k)$ for a specific k. Note from these panels that pervasive communities are extensively soft-overlapping. MDMC has only one parameter, α, which has turned out to be controlling the resolution of community detection (the smaller α, the network is decomposed into more communities of smaller sizes) [4, 5]. Therefore, we have calculated MaxRC as a function of α (Fig. 1b). KBN's SBM has no such resolution-controlling parameter and is required to predetermine the number of communities to which the network should be decomposed [8]. We therefore examined KBN's SBM for K =10, 20 and 30 (note that K =10 is consistent with the number of planted communities). MaxRC given by MDMC for a wide range of α surpasses that given by KBN's SBM for any K, indicating that MDMC outperforms KBN's SBM.

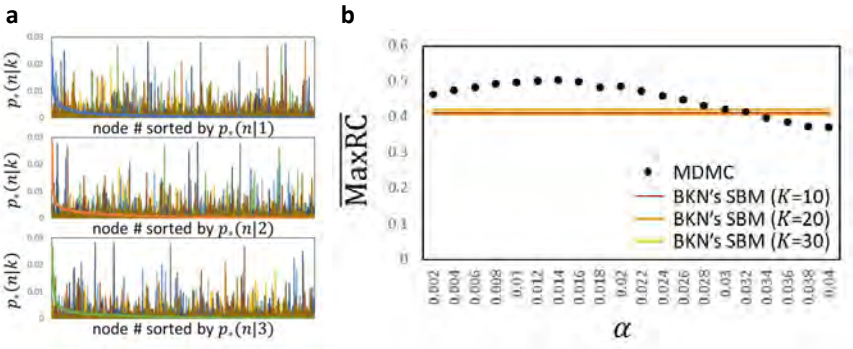


Fig. 1. (a) Pervasive structure of communities planted in the benchmark network. The same set of $\{p_*(n|k)\}_{k=1}^{K*=10}$ is shown in the top, middle and bottom panels with the node number (#) sorted in descending order of $p_*(n|1)$, $p_*(n|2)$ and $p_*(n|3)$, respectively. (b) MaxRC for MDMC averaged over 24 benchmark networks is plotted as a function of α (filled circle). MaxRC by BKN's SBM for K =10, 20 or 30 is indicated by red, orange or yellow horizontal line, respectively.

Computational cost of MDMC scales $\sim O(LK)$ with L being the total number of links, which means that it belongs to the fastest class of algorithms to detect pervasive communities [4, 5]. Together with this, the results obtained suggest that MDMC is a feasible approach to detecting pervasive communities from real-world networks. In the conference, we will demonstrate hierarchical organization of pervasive communities using brain networks constructed from connectome data of Allen Brain Atlas [9].

References

1. Lambiotte, R, Delvenne, J.C., Barahona, M. Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770v3 (2009)
2. Delvenne, J.C., Yaliraki, S.N., Barahona, M. Stability of graph communities across time scales. Proc Natl Acad Sci USA 107, 12755–12760 (2010)
3. Mucha, P.J. et al. Community structure in time-dependent, multiscale, and multiplex networks. Science 328, 876–878 (2010)
4. Okamoto, H., Qiu, X.-L. (2018). Community detection by modular decomposition of random walks. Complex Netwok 2018 (December 11-13, 2018, Cambridge, United Kingdom) Book Of Abstracts, 59–61
5. Okamoto, H., Qiu, X.-L. Modular decomposition of Markov chain: detecting hierarchical organization of pervasive communities. arXiv:1909.07066 (2019)
6. Page, L., Brin, S., Rajeev, M., Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, http://ilpubs.stanford.edu:8090/422/
7. Fortunato, S., Hric, D. Community detection in networks: A user guide. Phys Rep 659, 1–44 (2016)
8. Ball, B., Karrer, B., Newman, M.E.J. Efficient and principled method for detecting communities in networks. Phys Rev E 84, 036103 (2011)
9. http://connectivity.brain-map.org/

Not all Bridges Connect: Integration in Multi-Community Networks

Babak Heydari[1] and Pedram Heydari[2] Mohsen Mosleh[3]

[1] Northeastern University heydari@northeastern.edu
[2] Geisinger ADMI pheydari@geisinger.edu
[3] Massachusetts Institute of Technology, mmosleh@mit.edu

1 Introduction

There are many social and economic situations where two or more communities need to be integrated in an efficient and stable way that facilitates overall resource access throughout the network. We study structures for efficient integration of multi-community networks where building bridges across communities incur an additional link cost compared to links within a community. Building on the connections models with direct link cost and direct and indirect benefits, we show that the efficient structure for homogeneous cost and benefit parameters, and for communities of arbitrary size, always has a diameter no greater than 3. We further show that if the internal cost is not small enough to justify a full graph for each community, integration always follows one of these three structures: Single star, two hub-connected stars, and a new structure we introduce in this paper as *parallel hyperstar*, which is a special multi-core/periphery structure with parallel *bridges* that connect the core nodes of different communities and includes a wide range of efficiently integrated structures. Then we investigate stability conditions of these structures, using two different definitions: The standard pairwise stability, as well as a new stability notion we introduce in this paper as *post transfer pairwise stability*, which allows for bilateral utility transfers. We show that once post transfer pairwise stability is used, efficiency guarantees stability. Our results imply that both under and over integration (building too few or too many bridges) could negatively impact both stability and efficiency. More details of the results can be found in [1].

2 Model and Background Definitions

Agents, Networks, and Communities: Consider a set of nodes $\mathcal{N} = \{1, \ldots, n+n'\}$ each belonging to community I or I' (also called community 1 and 2 respectively) with $|I| = n$ and $|I'| = n'$. A network $\mathcal{G}$ is a set of pairs of agents $\{i, j\}$ that describes which agents are connected. We assume that the links are undirected and unweighted. For a given network $\mathcal{G}$, we use $N_i(\mathcal{G})$ to denote the neighborhood of node i,and $d_{ij}(\mathcal{G})$ to denote the distance (the minimum path length) between i and j.
Benefits, Costs and Utility: Following [2], the benefit that i receives from j is $b(d_{ij})$ for $b : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that for any $k > 0$, $b(k) \geq b(k+1) \geq 0$ and for any $k \geq n+n'$, $b(k) = 0$. Also, the cost of a link to j for i, denoted by c_{ij}, is c, if j is from the same community and is $c + \delta$, otherwise, for some $\delta \geq 0$.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

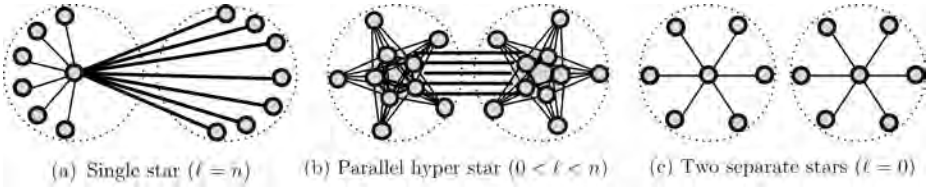


Fig. 1. Three efficient structures for two community networks.

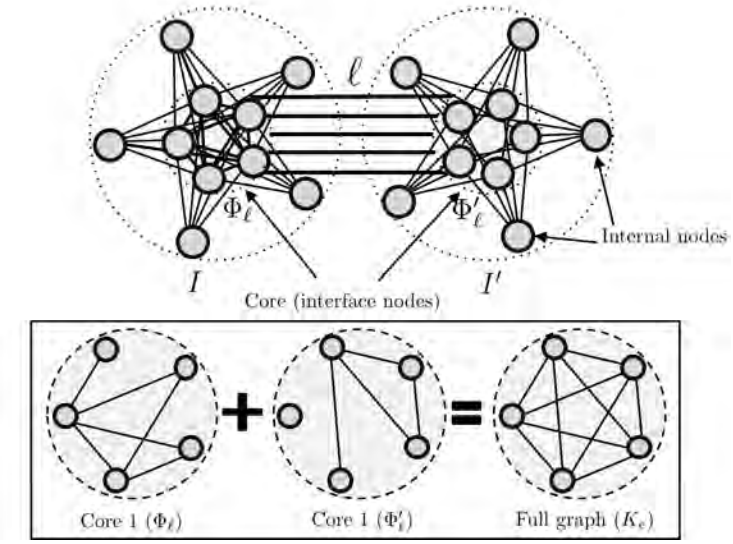


Fig. 2. Parallel hyperstar structure.

Let $u_i(\mathscr{G})$ and $U(\mathscr{G})$ represent the net utility that agent i receives under $\mathscr{G}$ and the total utility of $\mathscr{G}$, respectively. We assume $U(\emptyset) = 0$. Therefore, we have: $u_i(\mathscr{G}) = \sum_{j\in\mathscr{N}-\{i\}} b(d_{ij}(\mathscr{G})) - \sum_{j\in N_i(\mathscr{G})} c_{ij}$, and $U(\mathscr{G}) = \sum_{i=1}^{n} u_i(\mathscr{G})$.$\mathscr{G}$ is *efficient* if it maximizes $U(\mathscr{G})$. Also, $\mathscr{G}$ is (weakly) more efficient than $\mathscr{G}'$ if $U(\mathscr{G}) \geq U(\mathscr{G}')$. $\mathscr{G}$ is a *star* if there exists $i \in \mathscr{N}$ such that for any two distinct nodes $j,k \in \mathscr{N}$, $\{j,k\} \in \mathscr{G}$ if and only if $i \in \{j,k\}$. In addition to this standard structure, we introduce the following definitions.

Parallel Hyperstar Structure: One can regard a *parallel hyperstar* as a structure in which each community includes a *core*,a set of nodes that act as a *super-hub* where the connections across communities are only between those core nodes. An illustrative figure of the parallel hyperstar structure is depicted in Figure 2.

3 Efficient Integration

We focus on the less trivial (and more realistic) case where $c > b(1) - b(2)$ and a general benefit function. Our first result below shows that when integration does not make sense

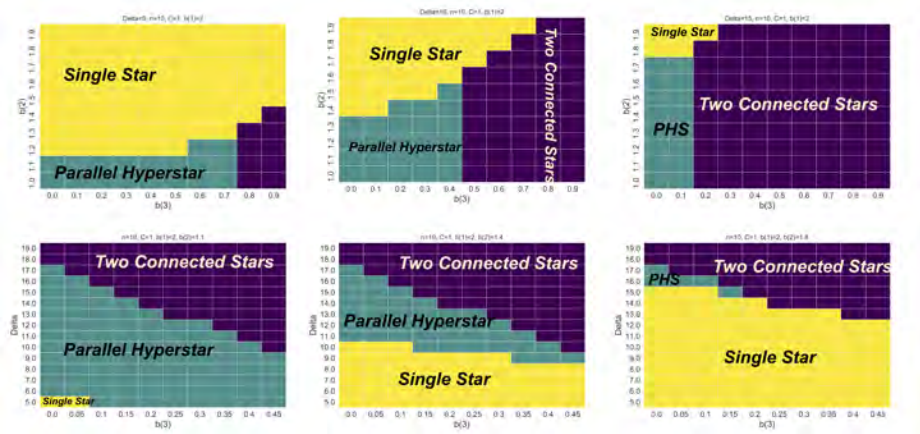


Fig. 3. *Most efficient structure (color coded) for each combination of model parameters. Only values that result in integration are shown.*

at all (i.e., it is not efficient to create any bridges between the two communities), then the efficient network is either empty, or it would consist of two separate stars, each residing in one of the two communities. On the other hand, when integration makes sense, then depending on the cost and benefit parameters only two classes for efficient structures are possible: Single star with a node from the larger community that acts as the global hub; and the new class of structures that we introduced in the previous section as parallel hyperstar. These results are formally stated in the following two theorems:

Theorem 1: *If the connection cost $c > b(1) - b(2)$, the efficient network is either an empty network, two separate stars, a parallel hyperstar, or a single star.*

In Theorem 1, we proved that the efficient network, if connected, is either a parallel hyperstar or a single star. The following theorem shows that when indirect benefits decay relatively slowly with distance, or the cost of forming internal links is relatively high, then a parallel hyperstar with more than 1 bridge cannot be efficient. The special case of 1 bridge is in fact a two (hub) connected starts structure.

Theorem 2: *When $c > b(1) - b(3)$, no parallel hyperstar with more than 1 bridge is efficient.*

Figure 3 shows sample plots where the efficient structure is color coded and labeled. When benefits fall considerably with distance, i.e., $b(d+1)$ is sufficiently smaller than $b(d)$, for a wide range of cross-community connection costs, parallel hyperstar is the most efficient structure. This makes parallel hyperstar a crucial design form for many practical applications where cost parameters in the mid-range, and benefits drop significantly as a function of distance.

4 Integration Stability

A central question of integration is that whether we can achieve simultaneous efficiency and stability, thus allowing a stronger role for direct intervention by a central authority during the integration process. We investigate the stability conditions of the efficient structures that were introduced in the previous section. We show that this is not in general possible, using the standard definition of pairwise stability. However, we argue that the standard definition of pairwise stability is too strict and does not include intuitive cases where agents are willing to subsidize formation or maintenance of some links through direct payment of cash or favour to their current or potential neighbors. To include such cases, we introduce a modified notion of stability, i.e. *pairwise stability with bilateral transfers* and prove that all three possible efficient structures are simultaneously stable.
A network $\mathscr{G}$ is pairwise stable, if for every two nodes i, j, the following two conditions hold:1) If $ij \in \mathscr{G}$, then $u_i(\mathscr{G}-ij) \leq u_i(\mathscr{G})$ and $u_j(\mathscr{G}-ij) \leq u_j(\mathscr{G})$. 2) If $ij \notin \mathscr{G}$ and $u_i(\mathscr{G}+ij) \geq u_i(\mathscr{G})$, then $u_j(\mathscr{G}+ij) < u_j(\mathscr{G})$. Based on this definition, we then show that simultaneous stability and efficiency isn't possible for integration structures in general.

Theorem 3: The single star structure is both efficient and stable when δ is small enough. A parallel hyperstar can be both stable and efficient only when it has just one bridge; i.e., it is a two-connected-stars.

Post Transfer Pairwise Stability*:* A network $\mathscr{G}$ is post transfer pairwise stable if for every $i, j \in \mathscr{N}$, we have: $1) ij \in \mathscr{G} \Rightarrow \Delta_i^{-ij}(\mathscr{G}) + \Delta_i^{-ij}(\mathscr{G}) \leq 0$, And, $2) ij \notin \mathscr{G} \Rightarrow \Delta_i^{+ij}(\mathscr{G}) + \Delta_i^{+ij}(\mathscr{G}) < 0$.Under this more realistic notion of stability, we can prove that efficiency can insure stability, in all the integration structures, particularly in the parallel hyperstar.

Theorem 4: Efficiency for parallel hyperstar structures guarantees post transfer pairwise stability.

This is a notable result, since it indicates that for all values of network parameters, the central authority can interfere in the integration process by leading the integrated network towards efficiency, and if pairwise direct transfer is allowed, the resulting network will automatically be stable. These results also indicates that both under and over integration (building too few or too many bridges) could negatively impact both stability and efficiency.

References

1. Heydari, B. Heydari, P., Mosleh, M., Not all Bridges Connect: Integration in Multi-Community Networks, Journal of Mathematical Sociology, forthcoming (2019)
2. Jackson, M., Wolinsky, A. :A strategic model of social and economic networks. J. of economic theory 71.1 44-74 (1996)

SOCS: A Fast Method for Overlapping Community Detection in Large Networks

Vinícius da F. Vieira[1], Carolina R. Xavier[1], and Alexandre G. Evsukoff[2]

[1] Department of Computer Science, Federal University of São João del-Rei, São João del-Rei, MG, Brazil,
vinicius@ufsj.edu.br,
[2] COPPE/Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

1 Introduction

One of the most important topological properties in complex networks is the organization of nodes as communities, a division of the nodes in groups with dense internal connections and sparse external connections. Different from traditional methods for community detection, which consider the division of the network as a partition problem, many works in the literature are focused on the identification of overlapping community structure in networks.

This work presents SOCS (Spectral Based Method for Overlapping Community Structure), a method for overlapping community detection built on the top of a traditional method for community partition: Newman's spectral method [5], a bisection divisive method for modularity optimization which calculates, at each bisection, the leading eigenvector of a modularity matrix, based on the adjacency matrix, solving a relaxed version of the modularity optimization problem. As proposed by Newman, after each bisection, the solution of the spectral method is improved by a fine-tuning stage, based on Kernighan-Lin algorithm [3] that swaps nodes from one community to the other in order to induce, at each step, the largest increase in modularity. SOCS performs a variation at the fine-tuning stage, allowing each node to belong to both communities in the bisection if it causes a positive gain to the modularity of both communities.

The proposed method is based on a high performance implementation of Newman's spectral method [8] and works with networks in the scale of millions of nodes, being able to be applied to several real world contexts. Preliminary experiments with real world benchmark networks, omitted in the current work due to lack of space, show that the method proposed in this work presents superior or similar quality when compared to state-of-art overlapping community detection methods.

2 Spectral Based Method for Overlapping Community Structure (SOCS)

In order to detect overlapping communities, SOCS performs successive bisections, as proposed by the original formulation of Newman's spectral method for modularity optimization. After each bisection, in [5], Newman proposes to swap nodes from communities in order to increase the modularity in a fine-tuning stage based on Kernighan-Lin

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

method. SOCS focuses on this fine-tuning stage to identify the overlapping nodes. As originally stated, Newman's original fine-tuning calculate which nodes were placed in a certain community (C_i) by the spectral stage but increases the overall modularity of division if placed in the other community (C_j). Then, Newman's spectral method swaps these nodes from C_i to C_j. The methodology proposed in this work is based on a slightly different idea: if a node contribute positively for the modularity of two communities C_i and C_j, then it should remain in both communities.

Some modifications are made to the original fine-tuning stage such that overlapping communities can be identified. The original fine-tuning stage evaluates, at each step, which are the nodes that may be swapped between two communities C_i and C_j in order to estimate those that cause the largest increase in the modularity when moved (from C_i to C_j or the opposite). The method stores the intermediate states and considers, as output, the one that represents the modularity with largest partition. For the overlapping approach, whenever a node increases the modularity in both C_i and C_j, it is considered as a member of both communities.

3 Results

After applying SOCS to a set of benchmark networks, the performance of SOCS can be assessed and compared to other methods found in the literature, regarding the execution time and the overlapping modularity Q_{ov} as proposed by Shen *et al.* [7]. The results are shown in Table 1. The computational environment consists of an Intel Core i9-9900K processor with 32Gb RAM running an Ubuntu 18.04 OS. The results presented

Table 1. Execution time (in seconds) and Overlapping modularity Q_{ov} for the studied networks.

	SOCS		CFinder [6]		Bigclam [10]		Demon [1]		COPRA [2]		SLPA [9]		OSLOM [4]	
	Time	Q_{ov}	Time	Q_{ov}	Time	Q_{ov}	Time	Q_{ov}	Time	Q_{ov}	Time	Q_{ov}	Time	Q_{ov}
CAHepPh[1]	1.64	**0.51**	-	-	5.03	0.35	74.65	0.14	2.37	0.16	4.31	0.25	480.86	0.46
CitHepTh[1]	8.65	**0.33**	-	-	35.37	0.16	86.18	0.04	9.67	0.01	15.02	0.14	1782.84	0.32
Dolphins[2]	0.01	**0.48**	0.01	0.29	0.45	0.08	0.12	0.28	0.09	0.26	0.08	0.39	0.34	0.37
Football[2]	0.01	0.54	0.02	0.55	1.89	0.16	0.15	0.27	0.10	0.25	0.15	0.24	0.36	**0.60**
Karate[2]	0.01	**0.40**	0.01	0.1147	0.27	0.09	0.11	0.04	0.10	0.18	0.07	0.33	0.10	0.367
Keys[2]	1.20	0.60	2111.34	0.38	1.93	0.63	2.98	0.31	1.64	0.68	2.50	**0.71**	122.20	0.38
Lesmis[2]	0.01	**0.53**	0.01	0.28	1.15	0.13	0.39	0.15	0.07	0.41	0.13	0.41	0.38	0.49
Polbooks[2]	0.01	0.48	0.02	0.43	1.84	0.13	0.15	0.08	0.13	0.35	0.16	0.43	0.56	**0.49**

in Table 1 allow us to see that SOCS is able to identify the overlapping community structure of the benchmark networks in a very low execution time. It is worth to notice that SOCS is able to identify communities in less than 10 seconds in networks with tens of thousands of nodes, as in the case of CitHepTh, with 27700 nodes and 352324 edges. SOCS can also identify overlapping communities in large scale networks (with more than one million nodes) in a reasonable time (less than one hour) without applying any sample strategy or reduction in the size of the network.

[1]Downloaded from: http://snap.stanford.edu/data/

[2]Downloaded from: http://www-personal.umich.edu/∼mejn/netdata/

Regarding modularity, community structure identified by SOCS presents a better quality than those identified by most other methods. For more than half of the networks explored, the communities detected by SOCS are more modular than those found by the other methods. For some networks, such as Karate and Les Miserables, the modularity measured for the communities extracted in this work is more than 25% higher than modularity observed for most competing methods. Even for networks where SOCS does not obtain the best results, Q_{ov} is very similar to the other results.

Despite the good results obtained in this work regarding modularity, other aspects can highlighted to confirm the suitability of the method on real world contexts. The method is based on a high performance implementation of Newman's spectral method [8], which allows it to be applied to large scale networks (in the order of million of nodes). The proposed methodology is simple, since it is based on a traditional and consolidated method for non-overlapping community detection, taking advantage of some benefits of the method, such as its well-known behavior.

For the next steps of the work, the resulting community structure identified by the proposed methodology should be more deeply studied in order to investigate the relationship between different network properties, such as modularity, community size and overlapping size, and better understand numerical aspects of the method. The adaptation proposed for the fine-tuning stage can be also applied to other non-overlapping community detection methods, potentially revealing high quality methods for community detection based on popular community detection methods. Yet, the methodology must be tested in a wider set of networks, in order to explore other contexts and, specially, larger scales of problems in a more appropriate computational environment.

References

1. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Uncovering hierarchical and overlapping communities with a local-first approach. ACM Trans. Knowl. Discov. Data 9(1) (2014)
2. Gregory, S.: Finding overlapping communities in networks by label propagation. New Journal of Physics 12(10), 103018 (2010)
3. Kernighan, B.W., Lin, S.: An Efficient Heuristic Procedure for Partitioning Graphs. The Bell system technical journal 49(1), 291–307 (1970)
4. Lancichinetti, A., Fortunato, S.: Limits of modularity maximization in community detection (Feb 2012), http://arxiv.org/abs/1107.1155
5. Newman, M.E.J.: Modularity and community structure in networks. PNAS 103(23), 8577–8582 (Jun 2006)
6. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043) (2005)
7. Shen, H.W.: Detecting the Overlapping and Hierarchical Community Structure in Networks, pp. 19–44. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
8. Vieira, V., Xavier, C., Ebecken, N., Evsukoff, A.: Performance evaluation of modularity based community detection algorithms in large scale networks. Mathematical Problems in Engineering 2014, 1–15 (December 2014)
9. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state of the art and comparative study. CoRR 1110.5813 (2011)
10. Yang, J., Leskovec, J.: Community-affiliation graph model for overlapping network community detection. In: Proceedings of the 2012 IEEE 12th International Conference on Data Mining. pp. 1170–1175. ICDM '12, IEEE Computer Society, Washington, DC, USA (2012)

Evaluating Nodes of Latent Mediators in Heterogeneous Communities

Hiroko Yamano[1], Kimitaka Asatani[2], and Ichiro Sakata[1,2]

[1] Institute for Future Initiatives, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan,
yamano@ifi.u-tokyo.ac.jp,
WWW home page: https://ifi.u-tokyo.ac.jp/
[2] Innovation Policy Research Center, Institute of Engineering Innovation, School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

1 Introduction

Knowledge heterogeneity has been investigated based on the observation of the benefits of integrating distant knowledge in the diversity of firm collaborations. Many researchers have demonstrated the effectiveness of incorporating knowledge from rare links, with widely accepted concepts such as shortcuts in *small world* [10], bridges between cliques as *weak ties* [5] and bridges over *structural holes* [1]. However, contrary to the prevailing conceptual works and case studies, there are fewer studies on the measurement of rareness of the links in a network.

The driving hypothesis of the present study is that the importance of a node is estimated from the heterogeneity of the links it brings. We already know, hubs which are nodes with many links, are important [4], but there is comparatively less evidence for the composition or values of the links that makes a node important. Most conventional network indexes, such as betweenness centrality [3], PageRank [8], or Burt's constraint [1] tend to be affected by the link density with adjacent nodes, which is sometimes unrelated to the community structure in the whole network. Although these indexes are effective to extract apparently significant nodes that have many important links [4], another method is required to find rare nodes that have a few important links.

In this paper, we propose an analyzing schema to comprehend the inter-community structure by combining the measures of nodal importance and community relevance. We demonstrate that the proposed index shows better performance compared to the participation coefficient P (P_i) in detecting nodes that connect distant communities. We validate the performance of the proposed index with the visualization of node rankings in networks with varied communities, and rank correlations, suggesting that proposed index identifies nodes that would make the average shortest path longer if they are if removed. Our approach sheds new light on node values by offering a way to detect latent mediators in heterogeneous communities with different number and density of nodes and links, that is consistent with the theories and numerous empirical studies in social and industrial networks [10, 5, 1].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

Firstly, we designed a new index PW_i by using experimentally verified community relevance (CRJC) and P_i. Based on the theory of information entropy [9], we took negative logarithm of the average of CRJC multiplied by P_i, and named the index as PW_i, which is defined as follows:

$$PW_i = -P_i \log \sum_{j \in \Gamma(i)^{IC}, j/=i} \frac{\mathrm{CRJC}(c_i, c_j)}{L+\delta} . \quad (1)$$

where P_i quantifies the proportion of links of the node i connecting to different modules [6]. CRJC(c_i, c_j) is Jaccard coefficient computed between the set of nodes and their neighbors of the communities c_i and c_j [2]. $\Gamma(i)^{IC}$ is the set of node i's neighbors that do not belong to c_i. L represents the number of the nodes in $\Gamma(i)^{IC}$. δ has an infinitesimal value of 0.000001 to prevent zero division error. We chose this equation because the amount of new information brought by node i to communities is given by the probability Pi of connecting communities while predicting the difference using existing knowledge, represented by community relevance.

Secondly, to estimate the performance of the new index, we generated the LFR network [7] with tightly connected 2 communities and weakly connected 3 communities. We visualized the network ranked by within-cluster degree Z [6], Katz centrality, betweenness centrality, P_i, PW_i, and inverse of Burt's constraint. P_i and PW_i ranked the nodes between communities highly, while the other indexes ranked the nodes within communities highly. In addition, only PW_i ranked the nodes mediating distant communities relationships higher than those connecting the most relevant communities (Fig.1).

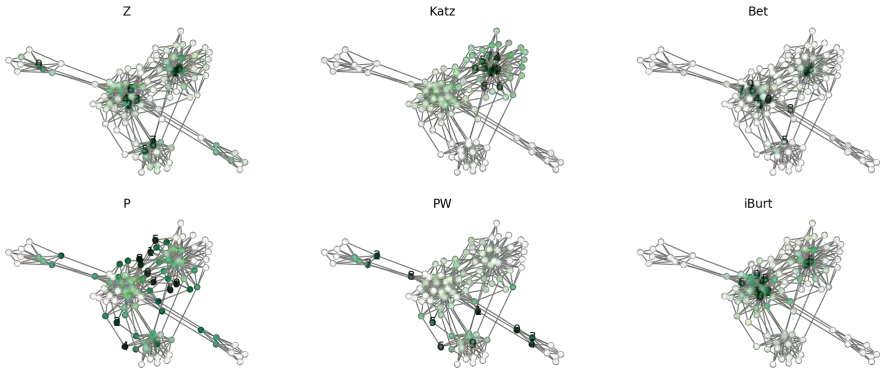


Fig. 1. Node rankings by six indexes in LFR network. Each network has 100 nodes colored by their values of corresponding index, 391 edges and 5 communities with 0.1 mixing rate. The labels of the ranking in each index are limited to top 10 nodes to avoid the complexity.

Thirdly, we investigated attack tolerance by measuring the average shortest path length (L) after removing a node sorted by community-based index of Z, P_i and PW_i.

Table 1. Rank correlation in LFR networks with varied link rates. Each network has 100 nodes, 8 degrees per node, 10 communities with 0.1 to 0.6 inter-community link rates.

mu	Z	P	PW
0.1	-0.424	0.617	**0.895**
0.2	-0.569	0.596	**0.838**
0.3	-0.435	0.549	**0.766**
0.4	-0.357	0.404	**0.504**
0.5	0.019	0.258	**0.280**
0.6	-0.021	0.082	**0.163**

We generated six LFR networks with different inter-community link ratio (mu), and calculated Spearman's rank correlation between the ranking ordered by each index and L. As a result, PW_i showed the highest correlations in all networks, suggesting that PW identifies nodes that if removed would make the average shortest path longer (Table1).

Summary. While node evaluation based on the adjacency relationship mainly uses local information, the community structure that characterizes the network has hardly been considered. In this study, we propose a new index that contributes to the understanding of the inter-community structure of a network by combining the measures of link distribution and community relevance. The visualization of node rankings and the rank correlations with respect to the attack tolerance of networks demonstrated that the proposed index showed the highest performance in comparison with five previously proposed indexes, suggesting a new way to detect latent mediators in heterogeneous networks.

References

1. Ronald S Burt. Structural holes and good ideas. *American journal of sociology*, 110(2):349–399, 2004.
2. Jingyi Ding, Licheng Jiao, Jianshe Wu, and Fang Liu. Prediction of missing links based on community relevance and ruler inference. *Knowledge-Based Systems*, 98:200–215, 2016.
3. Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.
4. Gourab Ghoshal and Albert-László Barabási. Ranking stability and super-stable nodes in complex networks. *Nature communications*, 2:394, 2011.
5. Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.
6. Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. *nature*, 433(7028):895, 2005.
7. Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.
8. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
9. Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.
10. Duncan J Watts. *Small worlds: the dynamics of networks between order and randomness*, volume 9. Princeton university press, 2004.

Part III

Diffusion and Epidemics

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

An extended SEIR model considering homepage effect for the information propagation of online social networks

Yu Guo, Danling Zhao, Jichao Li, Niping Jia, Jiang Jiang, and Yajie Dou

College of Systems Engineering, National University of Defense Technology, Changsha, Hunan, China, 410073
zhaodanling11@163.com

1 Introduction

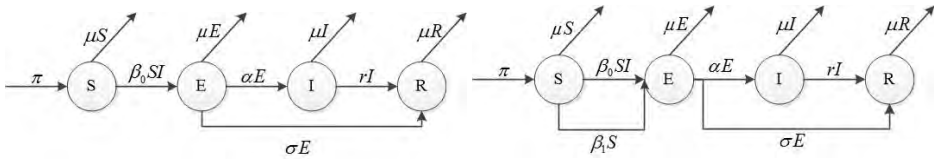
The Internet has changed modern lifestyle and played a role of reformation in information propagation. Nowadays information of all walks has occupied a large part of network users' daily routine, with various social media such as Facebook, Twitter and Wechat. Different information propagation platforms and spread ways determine the length of the informations survival time[1]. In the process of information propagation, a certain piece of information can be seen and received by the users in two ways. First, the information is pushed to the homepage with the users' support or the control of the websites manager, and then spread to all the online users. As most information never spread from the homepage, the second way is propagating among friends through the connecting networks. Researches on the propagation mechanism can discover the factors influenced the propagation rate, and then provide useful suggestions to control the spread process. Such researches have important applications for advertisers seeking, which expect to spread the advertisements to a range of net friends in a short time. In addition, a contrary application is to suppress public opinion, avoiding some incorrect information to spread fast and influence a large number of people. Here, the researches provide the mathematical models and theoretical analysis, so that people can make some targeted measures to control the propagation process.

2 Results

In the basic propagation model, parameters π and μ(indicating the rate that users enter and exit the OSNs respectively) are proposed to extend the original SEIR model, as shown in Fig.1(a). In the extended SEIR model, similar definitions from epidemiology[2] are used to categorize the users in information propagation. The susceptible population (S) consists of users who have not yet seen a certain piece of information, the exposed population (E) is made up of users who can see a certain piece of information because their connected users have forwarded it, the infected population (I) is composed of users who have forwarded a certain piece of information and it is visible by their connected users, the recovered population (R) is comprised of users who have forwarded a certain piece of information but (after a period of time) it is no longer visible on their followers homepages, or who have read but did not forward the information.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



(a) Basic propagation model (b) Propagation model considering homepage

Fig. 1. The propagation model

$$\cdot \begin{cases} S'(t) = \pi - \beta_0 S(t)I(t) - \mu S(t) \\ E'(t) = \beta_0 S(t)I(t) - \sigma E(t) - \alpha E(t) - \mu E(t) \\ I'(t) = \alpha E(t) - rI(t) - \mu I(t) \\ R'(t) = rI(t) + \sigma E(t) - \mu R(t) \end{cases} \tag{1}$$

The extended SEIR model can be presented as Eq. (1), where all the parameters are non-negative and defined as follows:π is the rate that users enroll the site; μ is the rate that users exit the site; β_0 is the transition rate from S to E ; σ is the transition rate from E to R; α is the transition rate from E to I; r is the transition rate from I to R. Since the explicit solution could not be found, the basic model reaches the equilibrium status[3] when the time-dependent ratios of $S(t)$, $E(t)$, $I(t)$, and $R(t)$ become constant. By calculation, two equilibrium points $E^1(S^*, E^*, I^*)$ and $E^2(S^{**}, E^{**}, I^{**})$ can be solved.

In the basic mode, the influence of homepage has not been taken into account.The Information on the homepage can be seen by all users in OSNs. Therefore, the information pushed onto the homepage will have a considerable influence because it faces to all users and have the maximum receivers. Assuming the probability of the users in OSNs who read the information on the homepage is β_1, which means there is another way for susceptible users transforming into exposed users. The propagation model considering homepage effect is shown in Fig.1(b).

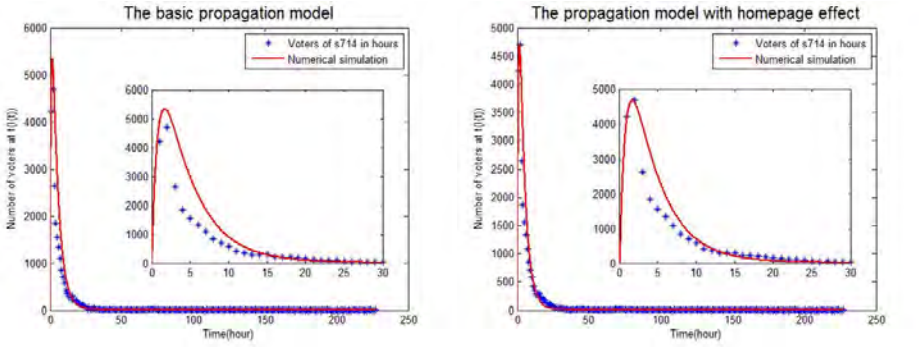
$$\begin{cases} S'(t) = \pi - \beta_0 S(t)I(t) - \beta_1 S(t) - \mu S(t) \\ E'(t) = \beta_0 S(t)I(t) + \beta_1 S(t) - (\sigma + \alpha + \mu)E(t) \\ I'(t) = \alpha E(t) - rI(t) - \mu I(t) \\ R'(t) = rI(t) + \sigma E(t) - \mu R(t) \end{cases} . \tag{2}$$

Similar with the solution in Eq. (1), the equilibrium point $E^3\left(\bar{I}, \bar{S}, \bar{E}\right)$ in Eq. (2) can be obtained.

To further verify the proposed models, the data of Digg.com is used for the case study.It is noted that the information diffusion mode in the Digg.com is good match with the extended SEIR model. Digg.com is an online social network platform where users are able to post content to a personal web page, vote for this content and share the content with other users who are connected with them. Once the posted content receives a large number of votes over a particular period of time, the content will be posted to the homepage of Digg.com and visible to all users. The dataset contains 3553 distinct stories (online content), the number of votes for a particular story, the particular users that voted for each story and the time at which each user cast the vote. It is noticeable

that this dataset only includes the stories which were promoted to the homepage of Digg.com in June 2009.

The most popular article, denoted by s714, is analyzed to display the one typical propagation mechanism. The actual data of I(t) and numerical simulation are shown in Fig 2.



(a) The basic propagation model (b) The propagation model with homepage effect

Fig. 2. The propagation model applied on Digg.com

At the end of each hour, the cumulative number of users who have voted for the story can be calculated and used as the metric of prediction. Herein, the predict accuracy is 82.97%, which is 11.79% lower than the accuracy of the propagation model with homepage.

Summary. In this paper, we have extended the SEIR model for investigating the information propagation in OSNs and obtained two equilibrium points with powerful proof. In addition, through introducing the homepage effect, a more complex and comprehensive model is proposed and only one equilibrium point is obtained. An important control parameter R_0, corresponding to the basic reproduction value in the infectious disease, has been constructed and analyzed. Finally, the paper has worked at the data of two articles in Digg.com, respectively representing two typical propagation mechanisms. The predictive accuracy is 94.76% for the one that posted on the homepage at the beginning and 94.27% for the second article, which has experienced basic propagation process and then pushed to the homepage. The results of case study verify the mathematical analysis and simulation experiments.

References

1. Yan Q, Wu L, Liu C, et al. Information propagation in online social network based on human dynamics[C]//Abstract and Applied Analysis. Hindawi, 2013, 2013.
2. Xu R, Wang Z, Zhang F. Global stability and Hopf bifurcations of an SEIR epidemiological model with logistic growth and time delay[J]. Applied Mathematics and Computation, 2015, 269: 332-342.
3. Nistal Riobello R, De la Sen Parte M, Alonso Quesada S, et al. On the Stability and Equilibrium Points of Multistaged SI (n) R Epidemic Models[J]. 2015.

The Probabilistic Backbone of Complex Correlation Networks

Catharina Elisabeth Graafland[1], José Manuel Gutiérrez[1], Juan Manuel López[1], Diego Pazó[1], and Miguel Angel Rodríguez[1]

Instituto de Física de Cantabria (CSIC-UC),
Avenida de los Castros s/n, 39005 Santander, Spain
catharina.graafland@unican.es

1 Introduction

In many practical applications networks are the natural representation of a complex system (airport connections, social networks). Often, however, the complex system only provides a general dataset (e.g. in climate) from which a *data-driven network* has to be constructed. A common approach for network construction is to establish links between nodes (variables, to be determined from the dataset) with pairwise correlation over a given threshold τ (Correlation Networks, CNs). In climate science the application of CNs have proven successful with a number of recent applications [1][2][3][4][5]. The choice of the threshold is however non-trivial and results in a trade-off between the statistical significance of the allowed connections and the richness of network structures unveiled [1][3]. In [7] we revisit CNs in the context of a climate application in which a variable X_i represents the monthly mean temperature in gridbox i (lattitude λ_i, longitude ϕ_i, $\Delta\lambda = \Delta\phi = 10^\circ$). We show that CNs by construction include redundant information in the network topologies. From a probabilistic perspective, this is expressed by over-parameterized probabilistic models when considering the underlying empirical Gaussian model with non-zero covariances for linked variables.

As an alternative approach to construct data-driven networks we propose the use of more sophisticated probabilistic Bayesian Networks (BNs), developed by the machine learning community as a data-driven modeling and prediction tool. A BN is learned by a structure learning algorithm that includes only the (pairwise and conditional) dependencies among the variables needed to explain the data (maximizing the likelihood of the underlying probabilistic model). The topology of a BN is much more sparse than the corresponding —in terms of similar likelihood of the data— CN, but allows to extract the same physical relationships when analyzed with complex network measures (clustered regions, communities, central nodes). Also, the probabilistic model (density function) obtained from the BN graph is parsimonious and contains only significant parameters making the model suitable for probabilistic inference. We therefore advocate the use of BNs instead of CNs to construct data-driven complex networks as they can be regarded, from both graph analytic and probabilistic perspective, as the *probabilistic backbone* of the underlying complex system.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

We analyzed complex CNs and BNs of increasing complexity (number of edges, $|E|$) considering different correlation thresholds τ, and different iterations of the structure learning algorithm [6], respectively. This resulted in CNs and BNs of sizes up to approximately 200000 and 8000, respectively. Global graph analysis, quantified by global graph measures as clustering coefficient and diameter, reveals that small CNs (in terms of $|E|$) capture local regions that are highly linked (e.g. the tropics and Antarctica), but only few long-distance links characterizing teleconnections. Distant teleconnected dependencies —resulting from large-scale atmospheric oscillation patterns— are in general weaker than local dependencies, but they are key for regional climate variability [8]. Bigger CNs do capture distant dependencies but show a high degree of redundancy in both local- and distant-link density. On the other hand, a small BN captures both locally clustered regions and long-distant dependencies without redundant links. The balance of local and large distance links plays a role when deeper topological analysis is to be performed. For example, a community division algorithm based on betweenness centrality can characterize a small BN in its most important (teleconnected) regions —Figure 1(a),— but struggles to characterize (small and big) CNs —Figure 1(b)— because of redundancy in the link distributions. Small CNs show community structures with many isolated regions that can not be grouped into a community and large CNs exhibit one giant community covering great part of the global area which is difficult to disassemble.

We also analyzed the networks from a probabilistic perspective, extending the networks to probabilistic models in which the edges in the network represent parameters in a Gaussian probability density function (pdf) —the global temperature dataset is assumed to be multivariate Gaussian—. Using cross-validation of likelihood values of the probabilistic models, optimum models were learnt with good generalization capabilities (avoiding overfitting): BN (1796 edges) and CN (3119 edges). Larger networks do explain the train dataset better but fail to explain the validation dataset, making physical features extracted of both topology and associated density function non-generalizable. We observe a discrepancy in the size of optimum CNs (around 3119 edges) and topological informative CNs; only CNs of size much bigger ($\sim$ between two and four times) than 3119 reveal an informative network topology. On the other hand, the size of a statistical optimal and topological informative BN coincide (Figure 1). The same conclusion can be drawn analyzing the associated probabilistic models of the networks on their capacity to propagate evidence (calculating conditional probabilities). Figure 1 shows the propagation of el Niño like evidence (significant alteration of temperatures in the Pacific Ocean: $E = 2$). Propagation of evidence in BN (1796) is on both local and global scale whereas in CN (3119) the propagation is only on local scale.

Summary. Correlation Networks (CNs) suffer from redundant information in their network topology. Bayesian Networks (BNs), on the other hand, include only non-redundant information (from a probabilistic perspective) resulting in a sparse topology from which generalizable physical features can be extracted. We advocate the use of BNs to construct *data-driven* complex networks as they can be regarded as the *probabilistic backbone* of the underlying complex system.

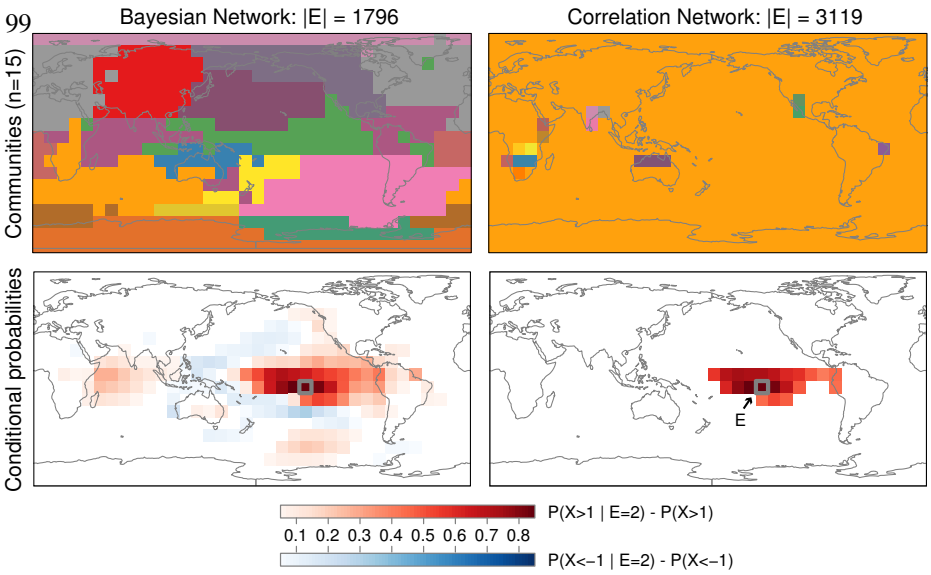


Fig. 1. First row: Results of community division algorithm when dividing the optimum BN (1796) and CN (3119) in 15 communities. The algorithm is not able to efficiently group the variables in communities for the CN. Second row: Propagation of El Niño like evidence (significant alteration of temperatures in the Pacific Ocean, $E = 2$) in optimum BN and CN probabilistic models. The maps show for each gridbox X the conditional probability of significantly increased (red color scale) or decreased (blue color scale) temperature given the evidence. The CN model only propagates the evidence on a local scale (i.e. does not capture teleconnections shown in the BN model).

References

1. Tsonis, A. A., Roebber, P. J.: The architecture of the climate network. Physica A: Statistical Mechanics and its Applications. 333, 497–504 (feb 2004)
2. Donges, J. F., Zou, Y., Marwan, N., Kurths, J.: Complex networks in climate dynamics. The European Physical Journal Special Topics. 174(1), 157–179 (2009)
3. Donges, J. F., Zou, Y., Marwan, N., Kurths, J.: The backbone of the climate network. Europhysics Letters. 87(4), 48007 (2009)
4. Boers, N., Rheinwalt, A., Bookhagen, B., Barbosa, H. M. J., Marwan, N., Marengo, J., Kurths, J.: The South American rainfall dipole: A complex network analysis of extreme events. Geophysical Research Letters. 41(20), 7397–7405 (2014)
5. Boers, N., Goswami, B., Rheinwalt, A., Bookhagen, B., Hoskins, B., Kurths, J.: Complex networks reveal global pattern of extreme-rainfall teleconnections. Nature. 566, 373-377 (2019)
6. Scutari, M., Graafland, C. E., Gutierrez, J. M.: Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms. International Journal of Approximate Reasoning. (To appear).
7. Graafland, C. E., et al.: The Probabilistic Backbone of Data-Driven Complex Networks: An example in Climate. (Submitted).
8. Yamasaki, K., Gozolchiani, A., Havlin, S.: Climate Networks around the Globe are Significantly Affected by El Niño. Phys. Rev. Lett. 100(22), 228501 (2008)

Distributed Epistemic Gossip Protocols*

Krzysztof R. Apt[1] and Dominik Wojtczak[2]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] University of Liverpool, Liverpool, UK

1 Introduction

Gossip protocols concern a set up in which each agent holds initially a secret and the aim it to arrive, by means of point-to-point communication (called *calls*) over a network, at a situation in which every agent knows all other secrets. During a call the agents involved exchange all secrets that they know. In other words, the aim of a gossip protocol is to generate a connected temporal graph [9].

Such protocols were successfully used in a number of domains, for instance communication networks [6], computation of aggregate information [8], and data replication [10]. For a more recent account see, e.g., [7].

In [4] a dynamic epistemic logic was introduced in which gossip protocols could be expressed as formulas. These protocols rely on agents' knowledge and are distributed, so they are distributed epistemic gossip protocols. This means that they can be seen as special cases of knowledge-based programs introduced in [5].

In [1] a simpler modal logic was introduced that is sufficient to define these protocols and to reason about their correctness. In [3] we showed that the distributed gossip protocols that use formulas of this logic are implementable and that their partial correctness and termination of these protocols is decidable.

In spite of this progress, several intriguing questions about distributed gossip protocols remain open. We discuss here these problems and establish some partial results.

2 Background

We assume a fixed set A of $n \geq 3$ ***agents*** each located on a node of a directed graph (digraph) and stipulate that each agent holds exactly one ***secret***. The secret of agent a is denoted by A, of agent b by B, etc. and the set of all secrets is denoted by Sec.

The language of our modal logic $\mathscr{L}$ is defined by the following grammar: $\phi ::= F_a S \mid \neg\phi \mid \phi \wedge \phi \mid K_a\phi$, where $S \in \mathsf{Sec}$ and $a \in \mathsf{A}$. So $F_a S$ is an atomic formula, while $K_a\phi$ is a compound formula. We read $F_a S$ as 'agent a is familiar with the secret S' (or 'agent a holds secret S') and $K_a\phi$ as 'agent a knows that formula ϕ is true'. Other Boolean connectives can be defined using $\neg$ and $\wedge$ in a standard way.

In the paper we shall use the following sublanguages of $\mathscr{L}$:

- $\mathscr{L}_0$, its propositional part, consists of the formulas that do not use the K_a modalities;
- $\mathscr{L}_1$ consists of the formulas without the nested use of the K_a modalities;
- $\mathscr{L}_1^a$, where $a \in \mathsf{A}$ is fixed, is a subset of $\mathscr{L}_1$ where the only modality used is K_a.

The goal of a distributed epistemic gossip protocol is to reach a gossip situation in which each agent is an *expert*, i.e., he knows all other secrets, starting at a gossip situation where each agent knows only his secret.

*This extended abstract is based on [2].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

In other words, their goal is to transform a gossip situation in which the formula $\bigwedge_{a\in\mathsf{A}}(F_aA\wedge\bigwedge_{b\in\mathsf{A},b\neq a}\neg F_aB)$ is true into one in which the formula $\bigwedge_{a,b\in\mathsf{A}}F_aB$ is true. Or, in the context of temporal graphs, the aim is to generate a connected temporal graph.

Let us recall the definition of the distributed gossip protocols [3]. By a ***component program*** for an agent a we mean a statement of the form $*[[]_{j=1}^{m}\ \psi_j\to \mathsf{c}_j]$, where $m\geq 0$ and each $\psi_j\to\mathsf{c}_j$ is such that a is the caller in the call c_j, and $\psi_j\in\mathscr{L}_1^a$ and all atomic formulas used in ψ start with F_a. If $m=0$, the component program is empty.

We call each such construct $\psi\to\mathsf{c}$ a ***rule*** and refer in this context to ψ as a ***guard***.

Intuitively, $*$ denotes a repeated execution of the rules, one at a time, where each time non-deterministically a rule is selected whose guard is true.

By a ***distributed epistemic gossip protocol***, from now on just a ***gossip protocol***, we mean a parallel composition of component programs, one for each agent. We call a gossip protocol ***propositional*** if all its guards are propositional, i.e., are from the language $\mathscr{L}_0$.

We presuppose that in each gossip protocol the agents are the nodes of a digraph and that each call ab is allowed only if $a\to b$ is an edge in the digraph. A minimal digraph that satisfies this assumption is uniquely determined by the syntax of the protocol. Let now us look at an example gossip protocol to which we shall return later.

Example 1. In [4] the following correct gossip protocol, called *Learn New Secrets* (LNS in short), for complete graphs was proposed. In the syntax of [1] used here, LNS is propositional, as it has the following component program for agent i: $*[[]_{j\in\mathsf{A}}\neg F_iJ\to ij]$. Informally, agent i calls agent j, if agent i is not familiar with j's secret.

Consider a gossip protocol P that is a parallel composition of the component programs $*[[]_{j=1}^{m_a}\ \psi_j^a\to\mathsf{c}_j^a]$, one for each agent $a\in\mathsf{A}$. By a ***computation*** of P we mean any call sequence $\mathbf{c}$ such:

- If $\mathbf{c}$ has finitely many calls then no guard ψ_j^a is true after all calls in $\mathbf{c}$ are made, i.e., $\mathbf{c}$ cannot be extended any further.
- For any prefix $\mathbf{c}'$ of $\mathbf{c}$, there exists a rule $\psi_j^a\to\mathsf{c}_j^a$ such that ψ_j^a is true after all calls in $\mathbf{c}'$ are made and $\mathbf{c}'.\mathsf{c}_j^a$ is also a prefix of $\mathbf{c}$. (Intuitively, this records the effect of the execution of the rule $\psi_j^a\to\mathsf{c}_j^a$ performed after the call sequence $\mathbf{c}'$ takes place.)

Any computation $\mathbf{c}$ corresponds naturally to a temporal (interval) graph. The k-th call in $\mathbf{c}$ where agent i calls agent j corresponds to an undirected edge from i to j with label $[k,\infty)$, i.e., this edge is active from the time point k onwards.

We say that the gossip protocol P is ***partially correct*** if for all its finite computations $\mathbf{c}$, after all calls in $\mathbf{c}$ are made, every agent is an expert. We say furthermore that P ***terminates*** if there are no infinitely long computations and say that P ***is correct*** if it is partially correct and it terminates.

3 Results

We begin with the following result for propositional gossip protocols.

Theorem 1 (cf. [2]). *Suppose that the agents form a star graph, so a graph in which some agent, say a, is present in all edges. Then no correct propositional gossip protocol exists for such a communication graph.*

Note that the LNS protocol from Example 1 shows that all complete digraphs have a correct propositional gossip protocol. We make here the following conjecture.

Conjecture 1. The class of graphs for which a correct propositional gossip protocol exists are digraphs with the property that the complement of the edge set does not contain a directed cycle.

One of the early results, see for instance [11], is that for $n \geq 4$ agents at least $2n-4$ phone calls are needed and sufficient to reach a situation in which each agent is an expert. However, such a gossip protocol is centralized and we conjecture here that it cannot be replicated in a distributed setting.

Conjecture 2. Prove that the lower bound $2n-4$ cannot be achieved for any distributed gossip protocol. In other words, prove that every correct gossip protocol for $n \geq 4$ agents generates computations of length $> 2n-4$.

We show that this conjecture is at least true for $n = 4$.

Theorem 2 (cf. [2]). *Every correct gossip protocol for 4 agents generates computations of length* > 4.

On the other hand, the following holds.

Theorem 3 (cf. [2]). *Suppose that* $n \geq 4$. *There exists a correct gossip protocol for* n *agents whose all computations are of length* $2n-3$.

We finally conjecture that this does not hold for propositional gossip protocols.

Conjecture 3. Prove that the lower bound $2n-3$ cannot be achieved by a correct propositional gossip protocol.

References

1. Apt, K.R., Grossi, D., van der Hoek, W.: Epistemic protocols for distributed gossiping. In: Proc. of TARK. EPTCS, vol. 215, pp. 51–66 (2016)
2. Apt, K.R., Wojtczak, D.: Open problems in the logic of gossips. In: Proc. of TARK (2017)
3. Apt, K.R., Wojtczak, D.: Verification of distributed epistemic gossip protocols. J. Artif. Intell. Res. (JAIR) 62, 101–132 (2018)
4. Attamah, M., Van Ditmarsch, H., Grossi, D., van der Hoek, W.: Knowledge and gossip. In: Proceedings of ECAI'14. pp. 21–26. IOS Press (2014)
5. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Knowledge-based programs. Distributed Computing 10(4), 199–225 (1997)
6. Hedetniemi, S.M., Hedetniemi, S.T., Liestman, A.L.: A survey of gossiping and broadcasting in communication networks. Networks 18(4), 319–349 (1988)
7. Hromkovič, J., Klasing, R., Pelc, A., Ruzicka, P., Unger, W.: Dissemination of Information in Communication Networks - Broadcasting, Gossiping, Leader Election, and Fault-Tolerance. Texts in Theoretical Computer Science. An EATCS Series, Springer (2005)
8. Kempe, D., Dobra, A., Gehrke, J.: Gossip-based computation of aggregate information. In: Proc. of FOCS. pp. 482–491. IEEE (2003)
9. Kempe, D., Kleinberg, J., Kumar, A.: Connectivity and inference problems for temporal networks. Journal of Computer and System Sciences 64(4), 820–842 (2002)
10. Ladin, R., Liskov, B., Shrira, L., Ghemawat, S.: Providing high availability using lazy replication. ACM Transactions on Computer Systems (TOCS) 10(4), 360–391 (1992)
11. Tijdeman, R.: On a telephone problem. Nieuw Archief voor Wiskunde 3(XIX), 188–192 (1971)

Modelling a Rehab-Recovery-Relapse Cycle

Iulia Martina Bulai[1,2], Benjamin Ortiz[3], and Andreia Sofia Teixeira[4,5]

[1] Department of Information Engineering, University of Padova, Italy
[2] Department of Mathematics, Informatics and Economics, University of Basilicata, Italy
[3] Accenture Federal Services, DC, USA
[4] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
[5] Center for Social and Biomedical Complexity, School of Informatics, Computing, & Engineering, Indiana University, Bloomington IN, USA
bulai@dei.unipd.it, benjamin.a.ortiz@accenturefederal.com, anmont@iu.edu

1 Introduction

In this work we explore the impact of community on individual behavior in a model we call the Rehab-Recovery-Relapse cycle. This model explores a system where individuals can either be susceptible to drug use or resistant to it. [1] In this system, the ability of individuals to change from one state to another is dependent on their community. That is, individuals can change if the community itself changes and becomes either more healthy or more susceptible. There is scholarship suggesting that the health of a community can be improved with the use of network intervention techniques.[2] These studies often involve training community members over the course of a day on how to motivate healthier practices among their peers. While these results are promising, they do raise concern with regards to the performance of drug rehabilitation centers. If individuals attending a short training can positively impact a community, then a similar positive effect should be seen from individuals who attend a rehab facility for a number of days. Rehab facilities have the benefit of removing a susceptible individual from a community. This provides individuals with time to recover in a healthier community as well as to learn how to positively affect their home community. However, unlike the network intervention studies, rehab facilities are in the open market and are not overseen by an individual researcher. Because of this, comparisons between network interventions and the rehab industry is not possible. In addition to this problem, it is also not easy to compare the communal impact of one rehab facility to that of another. The goal of the "Rehab-Recovery-Relapse" model is to provide a framework to examine the impact rehab facilities have on their communities. We hope to use this framework to not only compare rehab facilities with others, but to compare the practice of rehab facilities with other methods of drug intervention.

2 Methods

We consider a mathematical model that describes the behaviour of the population of a city composed by people with no addiction and that will not have it, people with addiction that can relapse, and then recovered in the community A or B. We assume that once

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the people recover they can became again addicted or not. The variables of the model are denoted by: H people with no addictions (and that will never have one); S people with addictions or that can become it; R_A people in the community of rehabilitation A; R_B people in the community of rehabilitation B. $N = H + S + R_A + R_B$ is the total human population. The mathematical model describes an hypothetical situation where the people can be distinguished well in four categories (people with no addiction, people with addiction, people recovered in community A and B, respectively). Notice that we assume to have two different rehab communities and three different environments, H and S live in the same place while R_A and R_B in other two different places (no relapse is considered). We introduce a model using ordinary differential equations. It means that every equation of the system give's us the behavior of the considered "population" in time. The model reads:

$$\begin{aligned}
\frac{dH}{dt} &= \Lambda + \phi\gamma_A R_A + \psi\gamma_B R_B - \mu_N H, \qquad (1)\\
\frac{dS}{dt} &= \Omega - \beta(H,S)S + (1-\phi)\gamma_A R_A + (1-\psi)\gamma_B R_B - \mu_N S,\\
\frac{dR_A}{dt} &= \delta\beta(H,S)S - \gamma_A R_A - \mu_N R_A,\\
\frac{dR_B}{dt} &= (1-\delta)\beta(H,S)S - \gamma_B R_B - \mu_N R_B, \quad \text{with} \quad \beta(H,S) = \frac{1+S}{1+H}.
\end{aligned}$$

First equation: describes the evolution of people without addictions. There is an immigration rate Λ of people with no addictions. While a part of the people that recover from community A (rate $\phi\gamma_A$) and/or B (rate $\psi\gamma_B$) can be strong enough to be introduced in this class. ϕ and ψ assume values in $[0,1]$, while γ_A is recovery rate in community A and γ_B in community B, respectively. We assume that people in class H dies naturally at rate μ_N. Second equation: we have the evolution of the addicted people, or those that can become it. Ω is the immigration rate of people with addictions. The people of this group can relapse at a rate $\beta(H,S)$. The proportion of the recovered people from A and B are $(1-\phi)$ and $(1-\psi)$, respectively. Third and fourth equations: are describing the populations in community A and B respectively. Once they relapse they are recovered in A or B, they recover at rates γ_A and γ_B, respectively. Both R_A and R_B dyes at a rate μ_N. $\delta \in [0,1]$. We assume that the parameters values are non-negative.

3 Results

The numerical simulations are made with Matlab. In particular we focus our attention on how the densities of the four populations at equilibrium change changing the values of two parameter values at the same time. Here we work with a rescaled version of model (1).

In Figure 1 are represented h, people without addictions, s, people with addictions, r_A, people recovered in community A, and r_B, people recovered in community B, respectively, at the equilibrium for values of per capita recovery rate in A, γ_A, and per capita recovery rate in B, γ_B, taking values in the intervals $[0,0.005] \times [0,0.005]$ and the remaining parameters values fixed. Notice that $\gamma_A = 0.005\ days^{-1}$ is equivalent to

saying that in community A it takes 200 days to recover. The right way to read two strain parameter plots is looking at the density of the considered population fixing a value of the parameter on x axes and see what happens when the value of the parameter on y axes is increased/decreased, and viceversa. In this way we know which role has the parameter values on the output of the system. We have done the two strain parameter analysis for (γ_A, δ), (γ_A, ϕ), (γ_A, ψ), (γ_A, δ) too, (results note reported here).

Summary. From Figure 1 we can conclude that for recovery rates 0.0015 $days^{-1}$ (667 days) the density of people in h remain constant at its maximum value, and the density of people in r_A and r_B at their minimum values, respectively, while for values smaller than this threshold, e.g. 0.001 $days^{-1}$ (1000 days) the density of people in h remain constant at its maximum value if the recovery rate of one community is the double of the recovery rate of the second community. This means that if the recovery rate it is not fast enough (assume values in $(0.001, 0.0015)$ $days^{-1} \simeq (1000, 667)$ days), the densities of the communities A and B increase with increasing γ_B and γ_A respectively, while for recovery rates 667 days, r_A and r_B remain constants. If the communities A and B collaborate to maintain a recovery rate higher than 0.0015 $days^{-1}$ then the density of people without addictions, h, will be at its maximum value.

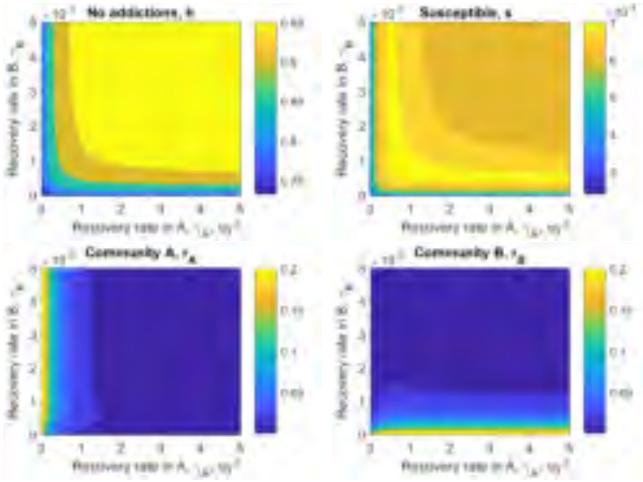


Fig. 1. On the first row: people without addictions (left), people that can have addictions (susceptible) (right); Second row: people recovered in community A (left), people recovered in community B (right); varying both γ_A and γ_B. Notice that the color scale is different in each panel.

References

1. Curtis, R., Friedman, A., Neaigus, B., Jose, B., Goldstein, M. and Ildefonso, G.: Street-level drug markets: network structure and HIV risk, Social Networks 17: 22949. (1995)
2. Latkin CA. Outreach in natural settings: the use of peer leaders for HIV prevention among injecting drug users' networks. Public Health Rep. 1998;113 Suppl 1(Suppl 1):151159.

Human prophylaxis driven by risk can cause oscillations in SIS like diseases

Benjamin Steinegger[1], Alex Arenas[1], Jesús Gómez-Gardeñes[2,3], and Clara Granell[2]

[1] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, E-43007 Tarragona, Spain,
benjaminfranzjosef.steinegger@urv.cat,
[2] Department of Condensed Matter Physics, University of Zaragoza, E-50009 Zaragoza, Spain
[3] GOTHAM Lab – Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, E-50018 Zaragoza, Spain

1 Introduction

Several infectious diseases display oscillations in the incidence through time. In a variety of cases, the subsequent outbreaks are caused by seasonal, exogenous events, such as the increase of influenza cases in winter, or the increase of vector-borne diseases during rainy seasons. However, there are diseases like syphilis which display non-seasonal periodic oscillations with a period of 8-11 years [1]. Different mathematical models aim to capture these non seasonal oscillations, either by considering models with temporary immunization [1, 2], or by allowing link rewiring in contact networks [3–5]. The aim of the latter is to incorporate the behavioral response of individuals, which eventually leads to sustained oscillations in the disease incidence.

In this work, we present a stochastic, yet analytically tractable, epidemic spreading model coupled with a two-strategy evolutionary game, which reflects the individuals decision on whether to take preventive measures. In this sense, agents can choose between two strategies *protected* (P) and *not protected* (NP). In general, the decision on prophylaxis is a trade off between the cost/effort of the prophylactic measures and their efficacy coupled with the severity of the disease. To describe this in a game theoretical framework, we introduce a protection cost c and an infection cost T. Additionally, agents have information about the global extent of the disease, which serves as an assessment of their infection risk. In this sense, we define the payoffs P_p and P_{np} associated to the two strategies P and NP as:

$$P_p = -c - T\frac{I_p}{S_p + I_p} \quad \text{and} \quad P_{np} = -T\frac{I_{np}}{S_{np} + I_{np}}\,. \tag{1}$$

The variables I_p and S_p represent the fraction of protected agents which are infected and susceptible, respectively. The same for I_{np} and S_{np}. Accordingly, the fractions $I_p/(S_p + I_p)$ and $I_{np}/(S_{np} + I_{np})$ describe the infection risk of a protected and not protected agent, respectively. In the temporal evolution, as the disease is spreading, agents adopt more successful strategies. We describe the disease spreading with an SIS model evolving on a synthetic network.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

As a first step, we will analyze the dynamics of our model with regard to the time averaged fraction of protected and infected individuals, which we show in Fig. 1 (a) and (b), respectively. We see that there are two critical values of the transmission probability, λ, in order to have a non zero fraction of protected individuals. For low values of λ, protection emerges as the disease is sufficiently infectious such that the reduced infection probability of protected individuals can actually compensate for the protection cost. Similarly, for high values of λ, protection vanishes as the quality of the prophylactic measures cannot balance the infectivity of the disease anymore. Furthermore, a mean field analysis of the system allows us to get an analytical approximation of the protection thresholds showing good agreement with the numerical solution. Regarding the epidemic incidence we observe that the epidemic threshold is not altered by the possibility of adopting prophylactic measures. As a matter of fact, at the epidemic threshold, the infection risk can still be considered zero. Accordingly, there is no incentive for individuals to adopt prophylactic measures and therefore the epidemic threshold is simply determined by the disease dynamics. In other words, the voluntary adoption of prophylactic measures allows to contain the disease but not to eradicate it.

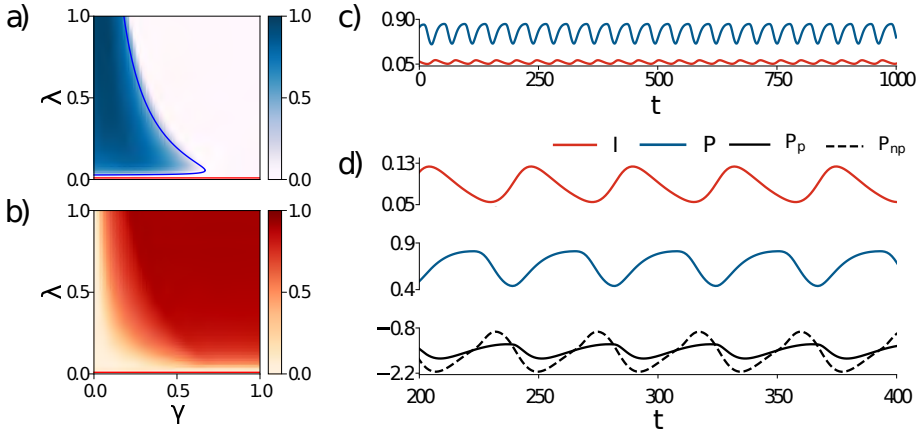


Fig. 1. Risk-driven epidemic spreading model. Numerical results of the risk–driven epidemic spreading model on a power-law network of size $N = 2000$ and exponent 2.5. Default parameters are $c = 1$, $\mu = 0.1$, $T = 10$. Phase-space diagrams for the transmission probability, λ, and protection quality, γ, of the incidence on the fraction of protected **(a)** and infected individuals **(b)**. Full protection is represented by $\gamma = 0$, while $\gamma = 1$ means that the prophylactic measures do not reduce the infection risk at all. The red line denotes the epidemic threshold of our model. The blue line is the protection threshold as obtained analytically from the mean field approximation. **(c)** Fraction of protected ($P = S_p + I_p$) and infected ($I = I_p + I_{np}$) individuals as a function of time. We observe an oscillatory behavior that is sustained in time. **(d)** Detail of the oscillations. The red and blue lines indicate the fraction of infected and protected individuals, respectively. The black dashed line plots the payoff of the strategy *not protected* (P_{np}) while the solid black line is the payoff of the *protected* strategy (P_p).

The second part of the results focuses on the temporal evolution of the system. Fig. 1(c) presents a trajectory of the system and we observe that the incidence of the epidemics, I, as well as the number of protected individuals, P, oscillates in time in a sustained way. In Fig. 1(d) we unveil the mechanism behind the oscillations: If the disease incidence is low, prophylactic measures are not beneficial and individuals stop adopting them. Therefore, the incidence increases before individuals eventually start adopting prophylactic measures again. At this point, the incidence will decrease and a new cycle can start. Additionally, we find that there is a critical value of the protection quality for oscillations to be sustained over time. If the protection quality is too low, the influence of the protection level on the epidemic incidence is not sufficient for having sustained oscillations. Instead, oscillations are damped and eventually vanish.
Finally, we propose plausible and efficient mechanisms to damp the oscillations. We show that targeted interventions, which are triggered as the disease incidence starts increasing, are much more effective than constant interventions of the same amplitude. In this sense, our study adds to the design of prevention campaigns, which do not only focus on perceived but real risks, in order to ameliorate human prophylactic behavior and contain future outbreaks as for example of sexually transmitted diseases.

In this work we present an analytically tractable epidemic spreading model in which individuals decide whether to take preventive measures or not depending on the global extent of the disease, being this an assessment of their infection risk. We show that the combined feedback between the human decision on prophylaxis, and the perceived epidemic risk, are sufficient conditions for the emergence of self-sustained oscillations in diseases well-described by the Susceptible-Infected-Susceptible (SIS) compartmental model. Finally, we propose plausible mechanisms to damp out the oscillations. Our study prompts to the design of persistent prevention campaigns, substantiated on not only perceived but real risks, to improve human prophylactic behavior and contain the recently reported raise of sexually transmitted diseases [6].

References

1. Grassly, N. C., Fraser, C. & Garnett, G. P. Host immunity and synchronized epidemics of syphilis across the united states. Nature 433, 417 EP (2005)
2. Hethcote, H., Stech, & Van Den Driessche, P. Nonlinear oscillations in epidemic models. SIAM Journal on Applied Mathematics 40, 19 (1981)
3. Gross, T., DLima, C. J. D. & Blasius, B. Epidemic dynamics on an adaptive network. Phys. Rev. Lett. 96, 208701 (2006)
4. Althouse, B. M. & He bert-Dufresne, L. Epidemic cycles driven by host behaviour. Journal of the Royal Society, Interface 11, 20140575 (2014)
5. Sherborne, N., Blyuss, & Kiss, I. Z. Bursting endemic bubbles in an adaptive network. Physical Review E 97, 042306 (2018).
6. Sexually transmitted infections and screening for chlamydia in England, 2018. Health Protection Report 13 (2019).

Degree dependent transmission rates in an epidemic model

Gareth Baxter and Gábor Timár

Departament of Physics & I3N, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
gjbaxter@ua.pt

1 Introduction

In [1], the authors argue that, in highly heterogeneous sexual contact networks, it is unrealistic to assume that the transmission risk per partnership is equal. Rather, an individual with numerous contacts does not transmit the infection to each partner with the same probability as an individual with few contacts. More generally, one would not expect infection transmission probabilities in an epidemic process to be uniform in a heterogeneous network.

Most studies of Susceptible-Infected-Recovered epidemic models on networks [2–4] assume that the infection begins with a vanishing fraction of the network, but not a finite number of sites. This essentially removes fluctuations with regard to the initial growth of the infection, and allows the problem to be mapped to a undirected bond-percolation on the same network [2]. This allows the calculation of the epidemic threshold (above which a finite fraction of the population is infected), expected epidemic size, and other statistics. In this construction, heterogeneity in infection probabilities has no effect, and subsequent works generally assumed a uniform infection rate when examining heterogeneous networks. Studies of epidemic models on networks have examined the effect of degree distributions and other network structure [3, 4] and neighbor degree correlations [5].

A few works [6–8], however, have considered epidemics originating with a single initial infection. In this case, one must consider not only the epidemic size but also the probability that it occurs. The bond percolation mapping is not sufficient, and instead a generalised directed percolation method must be used [7, 6]. The total expected epidemic outbreak is the product of two quantities: the probability that a single infection leades to a (giant) infection, and the expected size of the resulting epidemic (the probability that a random site receives the infection). These can be viewed as the giant IN- and giant OUT-components, respectively, of a specific directed network construction.

Here we use such an analysis to examine a compartmental epidemic model in which the transmission probability may depend on both the source and destination degrees. We consider a population of agents who may be in a susceptible (able to be infected), infected, or recovered (no longer infected and not able to be re-infected) state. Infected agents may pass the infection to susceptible neighbors, and the rate of transmission depends on the degree of both the infected and susceptible agents. We show that this heterogeneity can have a significant effect.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

We show that the epidemic threshold is strongly affected by degree-dependent transmission rate heterogeneity. Interestingly, the effect of dependence only on the source degree or only on the destination degree, is the same.

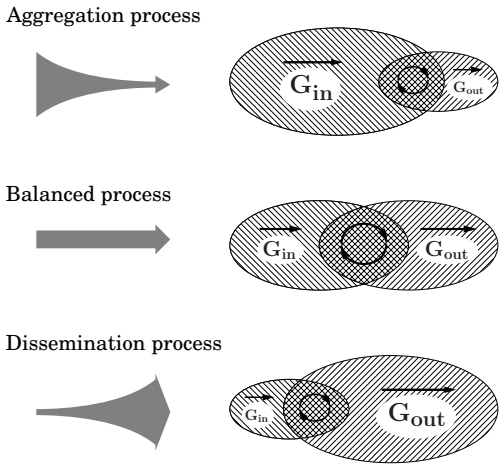


Fig. 1. Three different types of asymmetric spreading processes. The transmission probabilities and the degree distribution together determine what category a given process falls into. Symmetric processes (e.g., the standard SIR model) are perfectly balanced processes.

In the classical SIR model, the giant IN- and OUT- components are of equal size, but in the presence of heterogeneous transmission rates, they may be of different sizes, even when their product is the same. This has important implications: a infection that rarely produces a very large epidemic must be treated very differently to one that regularly produces a moderate epidemic. We therefore classify epidemics by the ratio of the IN- and OUT-component sizes (we quantify this by considering their ratio just above the epidemic threshold), Figure 1.

We give a general analysis of the problem for large locally tree like networks, when the transmission rate is an arbitrary function $\lambda f(k,k')$ of the source, k, and destination, k', degree (λ is used as a control parameter). We further give exact closed form solutions in the case of dependence only on source or on destination degree, and find approximate solutions in the case of dependence on both, valid for large mean degree or when the dependence on source and destination degrees is not far from symmetric.

We find a complex dependence of the process classification on the degree dependent transmission rate, which may be positively or negatively correlated with degree, Figure 2. Balanced, disseminating or aggregation processes variously occur according to the specific dependence.

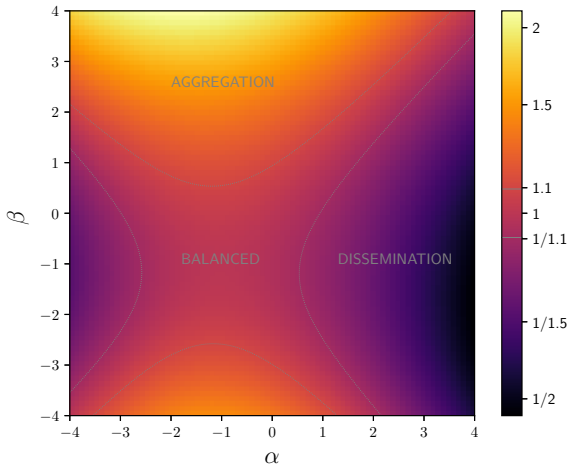


Fig. 2. Ratio G_{in}/G_{out} of the probability that a randomly selected site gives rise to an epidemic to the probability that a random selected node is infected in an epidemic (which gives the mean outbreak size), for transmission rates of the form $\lambda k_i^{\alpha} k_j^{\beta}$, where k_i is the degree of the infecting node, and k_j the degree of the potentially infected node.

Summary. We examine the effect of heterogeneous transmission rates, specifically rates depending on site degree, in a generalised SIR epidemic model on complex networks. We analyse the problem through a mapping to a generalised directed percolation problem. We classify processes as disseminating, aggregating or balanced, according to the ratio between the probability that a single infection leads to an epidemic and the probability that a site participates in the epidemic. We find a complex dependence on the transmission rate function.

References

1. Moslonka-Lefebvre, M., Bonhoeffer, S. and Alizon, S., Weighting for sex acts to understand the spread of STI on networks, J. Theor. Bio. 311, 46–53 (2012).
2. Newman, J., Spread of epidemic disease on networks, Phys. Rev. E 66, 016128 (2002).
3. Pastor-Satorras, R. and Vespignani, A., Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86, 3200 (2001).
4. Moreno, Y. and Pastor-Satorras, R., Epidemic outbreaks in complex heterogeneous network, Euro. Phus. J. B 26, 521–529 (2002).
5. Boguná, M. and Pastor-Satorras, R., Epidemic spreading in correlated complex networks, Phys. Rev. E 66, 047104, (2002).
6. Kenah, E. and Robins, J. M., Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing, J. Theor. Bio. 249, 706–722 (2007).
7. Miller, J. C., Epidemic size and probability in populations with heterogeneous infectivity and susceptibility, Phys. Rev. E 76, 010101 (2007).
8. Rogers, T., Assessing node risk and vulnerability in epidemics on network, EPL 109, 28005 (2015).

The effects of message sorting in the diffusion of low and high quality information in online social media

Diego F. M. Oliveira[1,2,3] and Kevin S. Chan[2]

[1] U.S Army Research Laboratory, 2800 Powder Mill Rd., Adelphi, MD 20783 USA,
[2] Network Science and Technology Center, Rensselaer Polytechnic Institute, Troy, NY - USA.
[3] diegofregolente@gmail.com, home page: diegofregolent.com

1 Introduction

The introduction of online social media platforms such as Twitter and Facebook, have changed completely the ways the modern civilization consumes and share information. If from one side they can facilitate the interaction between people from different parts of the globe, they also provide the perfect ground for the spreading of low-quality information such as fake news and misinformation (i.e., information that is misleading or inaccurate) that can be very harmful to our society. Traditionally, models of information diffusion are based on tools borrowed from theoretical epidemiology where susceptible agents became infected by interaction with infected agents and, in spite of their simplicity, they were able to reproduce several empirical observations. In situation in which quality is not easily quantified, other metrics - such as ratings, number of views, likes, number of downloads, etc - can be used the enhance the exposure of certain content to people. In principle, such an approach would allow high-quality information to prevail. However, such a popularity-based approach can create bias since the systems can be easily manipulated by social bots, for example. Another disadvantage of such approach was proposed by Sunstein and Pariser. They have argued that the reliance on personalization and social media can lead people to being exposed to a narrow set of point of views and one's existing beliefs would be reinforced because they are locked inside so-called filter bubbles or echo chambers, which prevent the users from engaging with ideas different from their own. Such selective exposure could facilitate confirmation bias and possibly create a fertile ground for polarization and misinformed opinions. Although several other works have been done trying to address to the crucial importance for the problem of competition for attention, there still a lack of a better understanding of how memes behave in on-line social network. In this work, we investigate how the way information is presented to the users will affect the system's quality, diversity and discriminative power. Here, we assume that each piece of information carries a numerical proxy representing its quality or truthfulness. We anticipate that by sorting the memes, we increase the exposure of high-quality information, therefore, increasing the overall system's quality. However, it is still unknown how it will affect other characteristics of the systems such as diversity of information and discriminative power.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

In this work we consider an agent-based model inspired by the long tradition of representing the spread of memes as an epidemic process. The model consists of a network where each agent is equipped with a memory containing α memes. Additionally, every meme has a quality represented by numerical value drawn from an uniform distribution. Furthermore, in contrast with classical epidemiological models, new memes are continuously introduced into the system in an exogenous way. We assume that at time $t = t_0$ the system is in its state of higher diversity where each node has α unique memes. At every time step a node i is selected at random and with probability μ it introduces a new meme in the system by adding it to its memory and sharing it with all its neighbors. On the other hand, with probability $1-\mu$ the selected node chooses a meme from its memory and, than, transmits it to all its neighbors. Once all neighbors receive the meme, we consider two situations, namely (a) the memes are organized as they are received in a first-in-first-out manner or (ii) in order to investigate the effects of quality bias, we assume that the user's memories are sorted according to the meme's popularity with more popular information on the top and less popular memes on the bottom of the node's lists. In both cases, the memes at the very bottom of the user's memories are removed or forgotten to make space for the incoming meme if the node does not have the meme already in its memory. Additionally, the probability that an agent selects a specific meme m from its list to transmit is proportional to the meme's quality $f(m)$ and it is giving by $P_i(k) = \frac{f(m_k)}{\sum_{j=1}^{\alpha} f_i(m_j)}$. Figure 1 shows the behaviour if the average system's quality as a function of time. Observe that, for long enough time, the system quality decreases significantly as the information load increases (Fig. 1 (a)). On the other hand, once sorting is introduce, high quality information prevails.

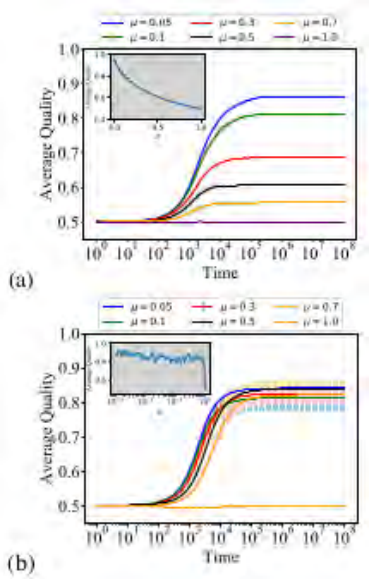


Fig. 1. Behaviour of the average quality as a function of time for the model (a) without sorting and (b) with sorting according to the meme's popularity. The insets show the behaviour of the average quality at the steady state for different values of the information load μ. The parameters used in all plots were and $\alpha = 14$.

Next, to measure the amount of diversity in the system at the steady state, we start from the entropy $H = -\sum_m P(m)\log P(m)$ where $P(m)$ is the portion of attention received by meme m, i.e., the fraction of messages with m across all of the user feeds. The sum runs over all memes present at a given time and is averaged over a long period after stationarity has been achieved.

Figure 2(a) shows the behavior of the diversity (system's entropy) for (a) the baseline model and (b) the models with sorted attention list according to the meme's popularity for different values of α and μ. Observe that the information load does not affect significantly the system's diversity in any significant way as shown in Fig. 2(c). On the other hand, as we will show next, the it does decreases considerably the system's ability to distinguish between memes. To measure the system's discriminative power, we employ the Kendall rank correlation between popularity and quality, which is computed by ranking memes according to the two criteria and then counting the number of meme pairs for which the two rankings are concordant or discordant, properly accounting for ties. The extreme case $\tau = 1$ indicates a perfect correlation between quality and popularity and fitter memes are more likely to go viral. On the other hand, if $\tau = -1$, the two rankings are completely discordant. Figure 2(a)and (b) show in color the Kendall correlation rank for the two models considered for different values of α and μ. We observed that in general the rank correlation decreases as the information load increases and a comparison between the models review that in reality the introduction of sorting in reality hinders the system's discriminative power with differences between models being as high as 82.5% as shown in Fig. 2(f) [1].

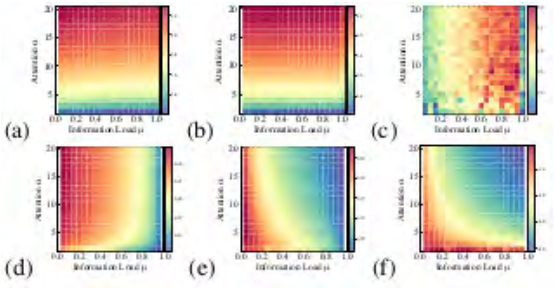


Fig. 2. The Diversity H (color scale bar) as a function of intensity of information load and attention for (a) the baseline model and (b) the model with sorted attention list according to the meme's popularity. The Kendall Tau(color scale bar) as a function of intensity of information load and attention for the (d) baseline model and the model with memes sorted according to (e) the meme's popularity. Figures (c) and (e) shows the difference in percentage between the two models.

Summary. We considered the problem of competition for limited attention. We investigated how message sorting affect the overall system's quality, diversity and discriminative power. Our results indicate that while the quality of information increases, the discriminative power decreases significantly. No significant change was observed for the diversity of information. We would like to thank the financial support by ARL through ARO Grant W911NF-16-1-0524. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Oliveira, D.F.M., Chan, K.S.: The effects of message sorting in the diffusion of low and high quality information in online social media, Under Review, (2019).

Google matrix of Bitcoin network: structure and contagion

Leonardo Ermann[1], Celestin Coquide[2] José Lages[2], and Dima Shepelyansky[3]

[1] Dto Física Teórica, GIyA, Comisión Nacional de Energía Atómica, Buenos Aires, Argentina
[2] Institut UTINAM, OSU THETA, Université. de Bourgogne Franche-Comté, CNRS, Besançon, France
[3] Laboratoire de Physique Théorique, IRSAMC, Université de Toulouse, CNRS,UPS, 31062 Toulouse, France

Abstract

We construct and study the Google matrix [4] of Bitcoin transactions during the time period from the very beginning in 2009 till April 2013 . Google matrix of the Bitcoin Network given by $G(\alpha) = \alpha S + (1-\alpha)\frac{1}{N} e e^T$ [1] built from data obtained from [2]

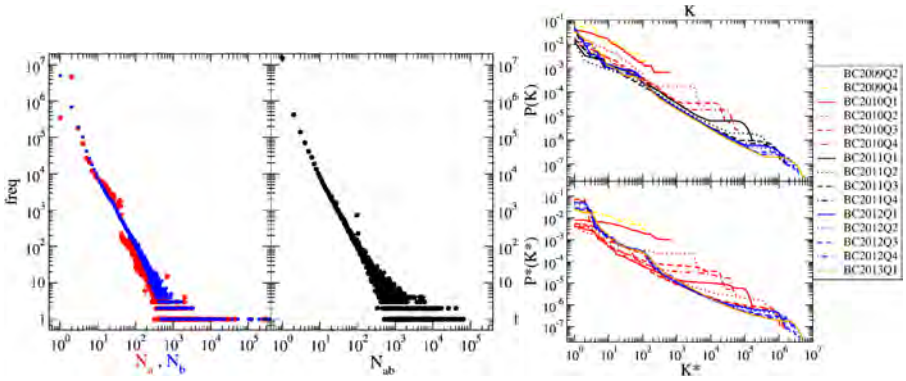


Fig. 1. Left panels show frequency histograms of Bitcoin Network (from January 11th 2009 to April 10th 2013) of transaction from user a (with red circles) to user b (with blue circles) and of given a to a given b (with black circles). Right panels show PageRank and CheiRank distributions ordered by indices K and K^* on top and bottom panel respectively. The bitcoin networks are taken by quarters of years (halfs in the case of 2009) for 2009 (yellow), 2010 (red), 2011 (black), 2012 (blue) and 2013 (orange) whith lines corresponding to Q1 (solid line), Q2 (dotted line), Q3 (dashed line) and Q4 (dot-dashed line).

The Bitcoin network has up to a few millions of bitcoin users and we present its main characteristics including some topology measures, the PageRank and CheiRank probability distributions, the spectrum of eigenvalues of Google matrix and related eigenvectors. We find that the spectrum has an unusual circle-type structure which we attribute to existing hidden communities of nodes linked between their members.

We show that the Gini coefficient of the transactions for the whole period is close to unity showing that the main part of wealth of the network is captured by a small fraction of users.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

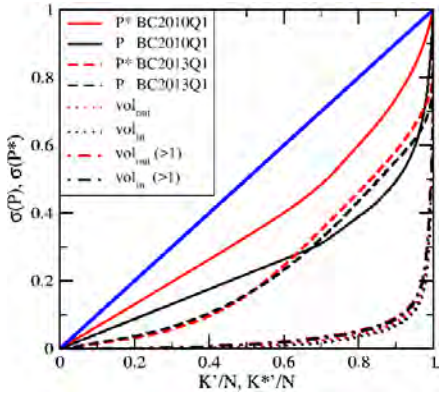


Fig. 2. Gini coefficient evolution for PageRank and CheiRank of BCN for quarter of years (halfs for 2009)..

We determine the dimensionless trade balance of each user and model the contagion propagation on the network assuming that a user goes bankrupt if its balance exceeds a certain dimensionless threshold κ.

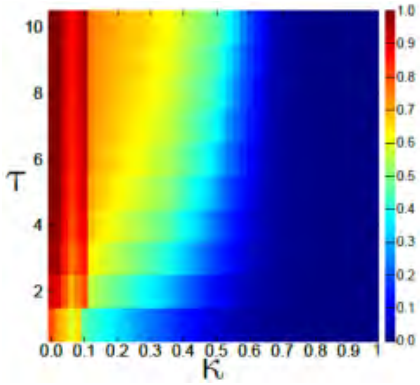


Fig. 3. Fraction N_u/N of BC13Q1 users in bankruptcy as a function of κ and τ.

We find that the phase transition takes place for $\kappa < \kappa_c \simeq 0.1$ with almost all users going bankrupt. For $\kappa > 0.55$ almost all users remain safe. We find that even on a distance from the critical threshold κ_c the top PageRank and CheiRank users, as a house of cards, rapidly drop to the bankruptcy. We attribute this effect to strong interconnections between these top users which we determine with the reduced Google matrix algorithm. This algorithm allows to establish efficiently the direct and indirect interactions between top PageRank users.

References

1. S. Brin and L Page, Computer Networks and ISDN Systems, 30, Issues 107 (1998),
2. https://blockchain.info/(accessed 25 Oct 2017)
3. Leonardo Ermann, Klaus M. Frahm and Dima L. Shepelyansky, *Google matrix of Bitcoin network*, Eur. Phys. J. B 91: 127 (2018).
4. L. Ermann, K.M. Frahm and D.L. Shepelyansky, *Google matrix analysis of directed networks*, Rev. Mod. Phys.87,1261 (2015).
5. Célestin Coquidé, José Lages, and Dima L. Shepelyansky, *Contagion in Bitcoin networks*, arXiv:1906.01293.

Effect of interaction bias on spreading dynamic in social networks

Matteo Neri[1] and János Kertész[1] Gerardo Iñiguez[1,2,3]

[1] Department of Network and Data Science, Central European University, H-1051 Budapest, Hungary,
Neri_Matteo@phd.ceu.edu, KerteszJ@ceu.edu, IniguezG@ceu.edu,
[2] Department of Computer Science, Aalto University School of Science, 00076 Alto, Finland
[3] IIMAS, Universidad Nacional Autonóma de México, 01000 Ciudad de Mexico, Mexico

1 Introduction

While some types of algorithmic biases have already been explored [1,2], a general framework to describe the effect of bias in social spreading is still lacking. We formalize the concept of bias in dynamical social systems in a general way by extending the well-known approximate master equation formalism [3,4,5,7,8,9]. In a stochastic binary-state dynamics, a node in the network can be in one of two possible states $[x(t) = 0, 1]$ at any time t and updates its status via infection and recovery rates ($F_{k,m}$, $R_{k,m}$) that depend on the degree k of the updating node and on the number of its infected neighbors m. The transition rates $F_{k,m}$ and $R_{k,m}$ fully characterize the temporal evolution of the node class (k,m). Our extended framework allows us to compute, in the presence of a bias with arbitrary functional form, effective transition rates both analytically and numerically, by means of approximations for several of the most studied binary dynamics of social spreading in networks (voter model, majority rule model, threshold model of complex contagion, etc.).

2 Results

As a concrete test, we implement algorithmic bias minimally to reflect the tailoring of information based on personal preferences on social networks. In order to filter the increasing amount of information produced on the web, one of the major biases introduced by online media platforms is a personalization of content according to the preferences of the user itself [6]. In order to do that, at the time of state-switching, we let a node disregard some of its neighbours in the opposite state with probability b (due to, e.g., an algorithmic bias to connect similar people in an online media platform): If $x = 0$, the node ignores $m - i$ of its m infected neighbours (each with probability b), and only considers i of them (each with probability $1 - b$).

We find that the introduction of bias in the selection process of interacting neighbors modifies the transition rates of several models of social spreading in non-trivial ways. The effective transition rates of a binary dynamics under the effect of bias are, instead,

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

expected values of the original transition rates over appropriate binomial distributions,

$$\begin{cases} F^*_{k,m} & = \langle F_{k-m+i,i} \rangle_{B_{m,i}(1-b)} \\ R^*_{k,m} & = \langle R_{m+s,m} \rangle_{B_{k-m,s}(1-b)} \end{cases}, \tag{1}$$

with $i = 0, \ldots, m$ and $s = 0, \ldots, k-m$ dummy variables over the number of infected/susceptible neighbours the node interacts with. We characterize the models according to the effect that the presence of bias induces on the dynamics, by observing how bias influences the time required to reach consensus, as well as by how bias amplification is related to the degree heterogeneity of the network or its mesoscopic (i.e. community) structure.

We observe that the combination of bias with sources of noise can induce new phases of behavior. In the case of the majority rule model [10], for example, noise indicates the probability that a user is not switching state even if the majority of its neighbors have the opposite opinion. The presence of algorithmic bias introduces a new phase of opinion polarization, as opposed to the known consensus phases where an initial majority dominates opinion (Figs. 1–2). Moreover, we investigate the microscopic effect of bias in inducing fragmentation and echo chambers in the system by observing how the auto-correlation and spatial correlation functions of the dynamics depend on bias.

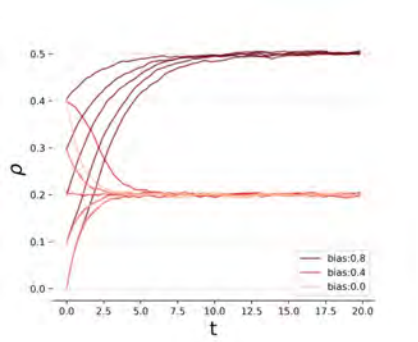


Fig. 1: Temporal evolution of the fraction of infected nodes $\rho(t)$ for an initially susceptible network with different initial conditions in the majority rule model with transition rates

$$F_{k,m} = \begin{cases} Q & \text{if } m < k/2 \\ 1/2 & \text{if } m = k/2 \\ 1-Q & \text{if } m > k/2 \end{cases}$$

and $R_{k,m} = 1 - F_{k,m}$. As the bias b increases, the system abandons the consensus equilibrium ($\rho = Q = 0.2$) with transient states of polarization that end up in a fully polarized network ($\rho = 1/2$).

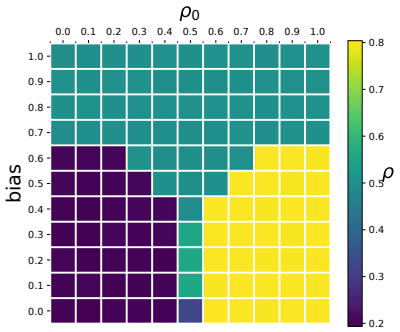


Fig. 2: Phase diagram of the fraction of infected nodes $\rho(t)$ (averaged over 50 realizations at Monte Carlo time $t = 60$) for the majority rule model over a regular random graph of size $N = 10^4$, as a function of bias b and initial condition ρ_0. For non-zero bias a phase transition appears, delineating a new regime of asymptotic opinion polarization (green), as opposed to the two known regimes of consensus (blue and yellow).

Summary. Our framework provides a principled way of exploring the generic effect algorithmic bias may have on any spreading dynamics in social networks. It shows that

some dynamics are robust against bias (notably, epidemic spreading and simple contagion such as the SIS and Bass models), while some others (opinion formation like the voter and majority rule models) show new phases of large-scale behaviour solely due to bias. By pinpointing common aspects among the diversity of biases and social interactions present in online environments, we identify idealized mechanisms to potentially tackle some of the most harmful effects of algorithmic bias, such as information bottlenecks, echo chambers, and opinion radicalization.

References

1. Sirbu., A. & Pedreschi, D. & Giannotti, F. & Kertész, J. : Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PLoS ONE* **14**, e0213246. (2019)
2. Perra, N. & Rocha: Modelling opinion dynamics in the age of algorithmic personalisation. *PLoS ONE* **9** 7261. (2019)
3. Gleeson, J.P.: High-accuracy approximation of binary-state dynamics on networks. *Phys. Rev. Lett.* **107**, 068701. (2011)
4. Gleeson, J.P. : Binary-State Dynamics on Complex Networks: Pair Approximation and Beyond. *Phys. Rev. X* **3**, 021004. (2013)
5. Porter, M. A. & Gleeson, J.P : Dynamical Systems on Networks. *Front. Appl. Dynam. Syst.* **4**. (2016)
6. Bozdag, E. : Bias in algorithmic filtering and personalization. *Ethic and Information Technology* **15**, 209-227 (2013)
7. Ruan, Z. & Iñiguez, G. & Karsai, M. & Kertész, J.: Kinetics of Social Contagion. *Phys. Rev. Lett.* **115**, 218702. (2015)
8. Karsai, M. & Iñiguez, G. & Kikas, R. & Kaski, K. & Kertész, J.: Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Sci. Rep.* **6**, 27178. (2016)
9. Unicomb, S. & Iñiguez, G. & Karsai, M. : Threshold driven contagion on weighted networks. *Sci. Rep.* **8**, 3094. (2018)
10. Krapivsky, P., & Redner, S.: Dynamics of Majority Rule in Two-State Interacting Spin Systems. *Phys. Rev. Lett.* **90**, 238701. (2003)

A dynamic contagion risk model with recovery features

Hamed Amini[1], Rui Chen[2], Andreea Minca[3], and Agnès Sulem[2]

[1] J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30303, USA, hamini@gsu.edu,
[2] INRIA Paris, 2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12, France
[3] Cornell University, School of Operations Research and Information Engineering, Ithaca, NY 14850, USA

1 Introduction

The random graph approach is a tool for systemic risk modeling when uncertainty stems from missing information on linkages. Such is the case for financial networks, see e.g. [1], [2]. Instead of who is connected to whom, only aggregated information at the level of each node is available. One can think of these as node characteristics, and examples include capital, asset or liability size, degree of connectivity. The random graph approach allows one to compute the limit (when the size of the network is large) of the fraction of nodes that fail when a shock propagates. The assumption is that one can categorize nodes according to some of their characteristics, and within each category, nodes are exchangeable. Along this direction, [3] assume that connectivity of each node is known and that the underlying graph is the configuration model, chosen uniformly over all graphs with the prescribed degree sequence. Their exchangeability assumptions on the linkage weights ensure that a limit exists for the fraction of nodes with an initial threshold to contagion. The final fraction of affected nodes is given in closed form for all values of degrees and initial thresholds.

Our main contribution in this paper is to extend the threshold contagion on the configuration model to the case when nodes' thresholds receive growth from the linkages. Because loss from the linkages and growth are intertwined, we call this *the recovery feature* of the threshold. We are motivated by the application to financial and insurance-reinsurance networks. Indeed, in financial networks thresholds represent –depending on the context – either capital or liquidity. An initial set of nodes fail exogenously and affect the nodes connected to them as they default on financial obligations. If those nodes' capital or liquidity is insufficient to absorb the losses, they will fail in turn. In other terms, if the number of failed neighbors reaches a node's threshold, then this node will fail as well, and so on. Since contagion takes time, there is the potential for the capital to recover before the next failure. It is therefore important to introduce a notion of growth.

The model we consider in this paper can be seen as a set of Cramér-Lundberg processes living on the nodes of a graph and which interact through the graph links. The capital grows linearly over time. In contrast to the Cramér-Lundberg process, losses do not arrive according to an exogenous Poisson process. Nodes have downward jumps when there is a failure of a neighboring node. When a node's capital or liquidity reaches zero, the node fails and it leads to downward jumps to its own neighbors. The notion

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

of time is also important. Calendar time governs the growth of capital. On the other hand, jumps are governed by the interaction between nodes (specifically between a failed one and one of its neighbors, chosen according to a probability law dictated by the random graph model). There is a natural notion of interaction time and the link revealing filtration. Consequently, jump arrival times have to be translated from interaction time to calendar time. We assume that inter-arrival times are exponentials with mean inversly proportional to the size of the network. We assume that in each time unit, nodes' growth is proportional to nodes' number of linkages. The linear growth as in the Cramér-Lundberg is also consistent with models in the wider network literature that attribute a fixed reward (respectively cost in some models) to each link as a tradeoff to more contagion risk (respectively network rewards), see [4], [5] and references therein.

2 Results

Let $\mu^{(n)}_{\lambda_+,\lambda_-,\theta}$ be the fraction of nodes with in-degree λ_+, out-degree λ_- and threshold θ. Assume the following regularity conditions $\mu^{(n)}_{\lambda_+,\lambda_-,\theta} \to \mu_{\lambda_+,\lambda_-,\theta}$, as $n \to \infty$, for some distribution $\mu : \mathbb{N}^3 \to [0,1]$. We also assume that the average connectivity converges to a finite limit

$$\bar{\lambda}^{(n)} := \sum_{\lambda_+,\lambda_-,\theta} \lambda_+ \mu^{(n)}_{\lambda_+,\lambda_-,\theta} = \sum_{\lambda_+,\lambda_-,\theta} \lambda_- \mu^{(n)}_{\lambda_+,\lambda_-,\theta} \to \sum_{\lambda_+,\lambda_-,\theta} \lambda_+ \mu_{\lambda_+,\lambda_-,\theta} =: \bar{\lambda} \in (0,\infty). \tag{1}$$

We assume that the duration in calendar time between the two successive interactions is given by a random variable $\Delta^{(n)}_k$ follows an exponential distribution of parameter n, i.e.,

$$\Delta^{(n)}_k = T^{(n)}_k - T^{(n)}_{k-1} \sim \mathrm{Exp}(n).$$

Suppose that growth benefits arrive uniformly over time according to the "growth parameter" α and both the in- and out-degrees. Given a growth function g, $g(\alpha,\lambda_+,\lambda_-)$, one can define the minimal time when the node could survive ℓ failed neighbors

$$t_\ell = t_{\lambda_+,\lambda_-,\theta,\ell} = \frac{(\ell-\theta)\bar{\lambda}}{g(\alpha,\lambda_+,\lambda_-)}. \tag{2}$$

Let $U^\pi_1, U^\pi_2, \ldots, U^\pi_\ell$ be i.i.d. uniform distribution on $[0,\pi]$ and the order statistics be

$$U^\pi_{(1)} \le U^\pi_{(2)} \le \cdots \le U^\pi_{(\ell)}.$$

Let us denote by

$$P_{\lambda,\theta,\ell}(\pi) := \mathbb{P}\left(U^\pi_{(\theta+1)} > t_{\theta+1}, U^\pi_{(\theta+2)} > t_{\theta+2}, \ldots, U^\pi_{(\ell)} > t_\ell\right), \tag{3}$$

for $\ell = \theta+1,\ldots,\lambda$ and $P_{\lambda,\theta,\ell}(\pi) = 1$ for $\ell = 0,1,\ldots,\theta$.

Theorem 1. *Let π^* be the relaxed fixed point of the map J^α defined as*

$$\pi^* := \min\{\pi \in [0,1] \mid J^\alpha(\pi) \le \pi\},$$

where

$$J^{\alpha}(\pi) := \sum_{\lambda_+,\lambda_-,\theta} \frac{\lambda_- \mu_{\lambda_+,\lambda_-,\theta}}{\bar{\lambda}} \cdot B^{\alpha}_{\lambda_+,\lambda_-,\theta}(\pi),$$

and

$$B^{\alpha}_{\lambda_+,\lambda_-,\theta}(\pi) := 1 - \sum_{\ell=0}^{\min\{\lceil \theta + g(\alpha,\lambda_+,\lambda_-)\pi \rceil - 1, \lambda_+\}} \binom{\lambda_+}{\ell} \pi^{\ell} (1-\pi)^{\lambda-\ell} P_{\lambda_+,\lambda_-,\theta,\ell}(\pi).$$

We have:

(i) If $\pi^* = 1$, i.e., if $J^{\alpha}(\pi) > \pi$ for all $\pi \in [0,1)$, then asymptotically (as $n \to \infty$) almost all nodes fail during the cascade.

(ii) If $\pi^* < 1$ and π^* is a stable fixed point of J^{α}, i.e., $J^{\alpha\prime}(\pi^*) < 1$, then the final fraction of failed nodes converges in probability to

$$\frac{|\mathscr{D}_f^{(n)}|}{n} \xrightarrow{p} \sum_{\lambda_+,\lambda_-,\theta} \mu_{\lambda_+,\lambda_-,\theta} B^{\alpha}_{\lambda_+,\theta}(\pi^*). \tag{4}$$

Our results show that a higher heterogeneity in the initial distribution of the threshold (as captured by its standard deviation) implies a lower default probability in equilibrium even as it leads to a larger average connectivity in equilibrium. More importantly, systems with higher growth/recovery rates can have equilibria with higher failure probability as well as higher final fraction of failed agents. The fact that bailouts lead to moral hazard problems is a known fact. Our results point to the fact that even in systems where threshold growth happens over time (as opposed to equity or liquidity infusions) strategic agents will adapt and potentially take more risks in equilibrium as captured by increased connectivity. This result is surprising. In anticipation of future growth agents take higher exposure to systemic risk and therefore the growth effect is hindered by higher exposures. To counteract this effect, the most interconnected agents should have higher thresholds in proportion to their interconnectedness, and this proportion should be even higher in environments with large growth. The effect of threshold growth over time would then allow them to play a role as shock absorbers in the system.

References

1. H. Elsinger, A. Lehar, and M. Summer: Risk assessment for banking systems. Management science, 52(9):1301–1314, 2006.
2. A. Gandy and L. A. M. Veraart: A bayesian methodology for systemic risk assessment in financial networks. Management Science, 63(12):4428–4446, 2017.
3. H. Amini, R. Cont, and A. Minca. Resilience to contagion in financial networks. Mathematical Finance, 26(2):329–365, 2016.
4. L. Blume, D. Easley, J. Kleinberg, R. Kleinberg, and E. Tardos: Network formation in the presence of contagious risk. In Proceedings of the 12th ACM conference on Electronic commerce, pages 1–10, 2011.
5. A. Majdandzic, B. Podobnik, S. V. Buldyrev, D. Y. Kenett, S. Havlin, and H. E. Stanley: Spontaneous recovery in dynamical networks. Nature Physics 10, 34, 2014.

Part IV

Dynamics on/of Networks

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Efficient limited time reachability estimation in temporal networks

Arash Badie Modiri[1], Márton Karsai[2,3], and Mikko Kivelä[1]

[1] Department of Computer Science, School of Science, Aalto University, FI-0007, Finland
[2] Department of Network and Data Science, Central European University, H-1051 Budapest, Hungary
[3] Univ Lyon, ENS de Lyon, Inria, CNRS, Université Claude Bernard Lyon 1, LIP, F-69342, LYON Cedex 07, France

1 Introduction

In many spreading processes a spreading agent may have a limited lifetime δt: like in case of transportation networks with a maximum acceptable transfer time; in social networks where information may become outdated or forgotten; or in case of diseases where the infectious period ends after a certain amount of time. These problems, concerning limited ($\delta t < \infty$) waiting time processes, have been previously studied in temporal networks by simulating the process from a sample of initial nodes and time instances. This approach limits the analysis to either very small networks, or average statistics (as opposed to event-level statistics or statistics of the tails of distributions) [3]. To alleviate this problem, recently the *event graph* representation has been proposed [4, 6], with weakly connected components giving an upper bound on the number of events (resp. nodes) what a spreading process can follow (resp. reach) [4]. However, as weakly connected components of event graphs cannot determine the exact reachable set from a node at a given time, the detection of out-components appeared as an open challenge so far.

In this contribution, we present a set of algorithms based on probabilistic cardinality estimation [1, 2] that allows us to simultaneously measure the number of nodes and events that can be reached from all different starting points and times in a temporal network. In its most basic form it consists of scanning through each node of the event graph (corresponding to events of the temporal network) in reverse topological order and constructing an out-component set for each node based on its successors.

Our work has several advantages as compared to the conventional initial condition sampling approach. It can be used to accurately calculate the tails of the reachability and spreading distributions and it can answer completely new questions on temporal network data, such as, what is the exact maximum number of nodes that can be infected via a spreading process. It can also be used to calculate node/event level statistics, which may lead to new kinds of importance and centrality measures. Further, it opens up a way to analyse percolation phenomena in temporal networks. For example, instead of resorting to upper-bounds via weakly connected components calculations (and lower bounds via sampling), we can now exactly measure the critical parameters of the temporal network unfolding as a directed percolation, or a spreading process evolving on the top of it.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Further, our method can find the event, which reaches the largest fraction of the network (the largest out-component in the event graph) with high adjustable probability. Note that the reachability without limited waiting time ($\delta t = \infty$) appears as a special case here and can be solved as well with our algorithms.

2 Results

Our method works accurately for very large networks, which we demonstrate via the estimation of reachable set of nodes and events from all possible initial conditions in a large mobile phone call network with $\sim$ 325M events [4] and a Twitter mention network of $\sim$ 258M interactions [7]. It can also be applied to directed temporal networks and networks with a delay between the start of the event and the time it takes effect. To demonstrate this, we applied the same method to the public transportation network of Helsinki with $\sim$ 664K events [5] and air transportation network of the United States of America with $\sim$ 180K events. Fig. 1 and Table 2 compare results and runtime of the estimation algorithm on the real-world networks mentioned above.

Table 1. Runtime for real-world networks when calculating the reachability (number of unique reachable events, nodes and lifetime) from all events in the network. δt^* corresponds to a waiting time around the time at which there is a jump in the largest out-component size and corresponds to the grey vertical line in Fig. 1. Baseline algorithm scans events in order of time and marks each event/node that would be reachable from a specific starting event. This is repeated for each event in the network as the starting event.

			Runtime		**Baseline**	
Name	**Events**	**Error**	$\delta t = \infty$	$\delta t = \delta t^*$	$\delta t = \infty$	$\delta t = \delta t^*$
Mobile	325M	3.3%	106 minutes	85 minutes	1695 years	21 years
Twitter	258M	3.3%	90 minutes	77 minutes	2409 years	243 years
Public transport	664K	0.81%	59 minutes	60 seconds	19 hours	13 minutes
Air transport	180K	0.81%	235 minutes	17 seconds	138 minutes	60 seconds

References

1. Flajolet, P., Fusy, É., Gandouet, O., Meunier, F.: HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In: Jacquet, P. (ed.) AofA: Analysis of Algorithms. DMTCS Proceedings, vol. DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07), pp. 137–156. Discrete Mathematics and Theoretical Computer Science, Juan les Pins, France (Jun 2007), https://hal.inria.fr/hal-00406166
2. Heule, S., Nunkesser, M., Hall, A.: Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In: Proceedings of the 16th International Conference on Extending Database Technology. pp. 683–692. EDBT '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2452376.2452456
3. Holme, P.: Network reachability of real-world contact sequences. Phys. Rev. E 71, 046119 (2005), https://doi.org/10.1103/PhysRevE.71.046119

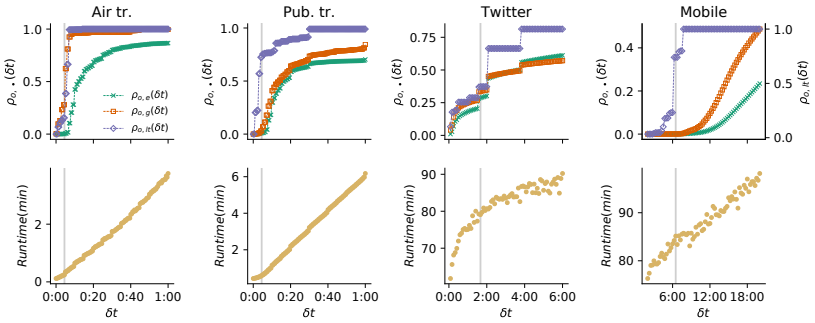


Fig. 1. Maximum out-component sizes (top row) based on number of events ($\rho_{o,e}$) number of unique nodes ($\rho_{o,g}$) and lifetime of the out-component ($\rho_{o,lt}$) and corresponding median runtime (bottom row) for different value of δt. The vertical line in each plot corresponds to the δt^* value in Table 2.

4. Kivelä, M., Cambe, J., Saramäki, J., Karsai, M.: Mapping temporal-network percolation to weighted, static event graphs. Scientific reports 8(1), 12357 (2018), https://doi.org/10.1038/s41598-018-29577-2
5. Kujala, R., Weckström, C., Darst, R.K., Mladenović, M.N., Saramäki, J.: A collection of public transport network data sets for 25 cities. Scientific data 5, 180089 (2018), https://doi.org/10.1038/sdata.2018.89
6. Mellor, A.: The temporal event graph. Journal of Complex Networks 6(4), 639–659 (2017), https://doi.org/10.1093/comnet/cnx048
7. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 177–186. ACM (2011), https://doi.org/10.1145/1935826.1935863

On consensus over heterogeneous temporal networks

Lorenzo Zino[1,2], Alessandro Rizzo[3,4], and Maurizio Porfiri[1,5]

[1] Department of Mechanical and Aerospace Engineering, New York University Tandon School of Engineering, Brooklyn NY, US
mporfiri@nyu.edu,
[2] Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands
lorenzo.zino@rug.nl,
[3] Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
alessandro.rizzo@polito.it,
[4] Office of Innovation, New York University Tandon School of Engineering, Brooklyn NY, US,
[5] Department of Biomedical Engineering, New York University Tandon School of Engineering, Brooklyn NY, US

The consensus problem is defined as a network of dynamical systems which coordinate toward a common state following a distributed algorithm. In view of its broad range of applications, encompassing opinion formation, distributed estimation, and multi-vehicle coordination, the consensus problem has received an ample attention in the last decades [1, 2]. However, most of the literature focuses on static networks, challenging our understanding of phenomena that are typically modeled by time-varying networks [3].

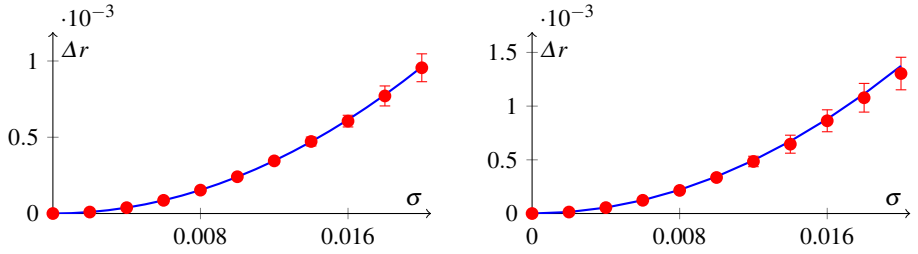
Here, we study the discrete-time consensus problem over time-varying, stochastic networks, by using the activity driven network (ADN) modeling framework [4, 5]. Within the paradigm of ADNs, each node is characterized by a fixed parameter, called activity potential, which encapsulates its propensity to communicate and exchange information with its peers. In plain worlds, the activity potential measures the probability that a node is activated in a time unit. The distribution of the activity potentials across the nodes models heterogeneity in individuals' behavior. ADNs are a powerful tool to study dynamical systems on networks. In fact, *i*) they allow for representing networks with a desired level of heterogeneity in the nodes' propensity do generate connections, in contrast with existing models of time-varying, stochastic networks [6], and *ii*) they beget mathematical models that are analytically tractable and amenable to fast simulations [4, 5].

Some preliminary endeavors toward a mathematical treatment of consensus problems over ADNs can be found in [7, 8]. Therein, results are mostly based on numerical simulations and on the assumption of a time-scale separation between the evolution of the network and the nodes' dynamics. Here, we build on these first endeavors toward a rigorous treatment of consensus over ADNs. Our technical contributions are twofold: *i*) we study mean-square convergence of the dynamical process to estimate the speed of convergence of the self-coordination process, and *ii*) we characterize the consensus state, that is, the expected common state reached by the dynamical systems [9].

To achieve the first result, we leverage methods from stochastic stability theory and we utilize a second-order eigenvalue perturbation argument. Specifically, building on the claims in [6], we derive closed-form results for the rate of convergence of the mean-square error dynamics as a function of the model parameters. We establish that the

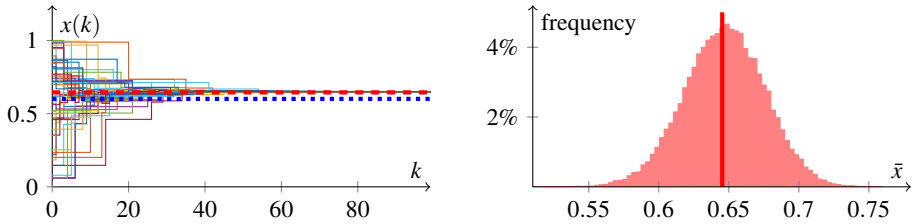
COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



(a) 30 nodes (b) 100 nodes

Fig. 1. Variation of the convergence factor with respect to the case of nodes with homogeneous activity, for increasing values of standard deviation of the activity distribution (denoted by σ), for two different choices of the model parameters with increasing network sizes. We observe that the numerical estimations performed over 100 independent runs (red circles, error bars are 95% confidence intervals) confirm our analytical prediction (blue curve).



(a) Sample path (b) Estimated consensus state

Fig. 2. Numerical simulations of the consensus dynamics on a network with 50 nodes. Activity potentials are distributed according to a power-law. Panel (a) illustrates a sample path of the process and compares the evolution of the state variables with the predicted consensus state (red dashed line) and the average of the initial conditions (blue dotted line). Panel (b) illustrates the empirical distribution of the consensus values for set of Monte Carlo simulations over 50,000 independent runs from the same initial condition of the state variables. The distribution seems to be centered in correspondence of our analytical prediction (red line).

convergence factor increases with the square of the standard deviation of the activity distribution. The larger is the convergence factor, the slower is the convergence to the consensus state. Hence, we suggest that the speed of convergence could be hindered by the heterogeneity of the nodes' activities, at least for moderate levels of heterogeneity. Figure 1 illustrates the results of a campaign of Monte Carlo numerical simulations, which confirms our analytical predictions.

The second result is attained using stochastic stability theory, whereby we characterize the expected value of the consensus state reached by the network nodes. Different from homogeneous systems, where the expected consensus state coincides with the average of the initial conditions, our analytical findings lead us to conclude that the consensus state is dominated by low-activity nodes. Figure 2 shows numerical simulations of the evolution of the network of dynamical systems, supporting our analytical results.

Toward the application of our modeling framework in real-world large-scale problems, we derive a set of asymptotic results in the limit of large networks, both for the rate of convergence and the consensus state.

Finally, we discuss the scenario where some of the network nodes act as leaders, steering the state of the whole network to their own state. Utilizing a first-order eigenvalue perturbation argument, we show that, in the presence of leaders, heterogeneity among the nodes could be beneficial to group decision-making. In fact, in [10] we prove that moderate levels of heterogeneity decrease the convergence factor, speeding up the convergence process to consensus.

Funding Statement

This work was partially supported by the National Science Foundation under grant No. CMMI-1561134; Compagnia di San Paolo, Torino, Italy; and the Italian Ministry of Foreign Affairs and International Cooperation, within the project Mac2Mic.

References

1. Ren, W., Beard, R.: Distributed Consensus in Multi-vehicle Cooperative Control. 1 edn. Springer Verlag, London, UK (2008)
2. Olfati-Saber, R., Fax, J.A., Murray, R.: Consensus and cooperation in networked multi-agent systems. Proc. IEEE **95**(1) (2007) 215–233
3. Cao, Y., Yu, W., Ren, W., Chen, G.: An overview of recent progress in the study of distributed multi-agent coordination. IEEE Trans Ind. Informat. **9**(1) (2013) 427–438
4. Perra, N., Gonçalves, B., Pastor-Satorras, R., Vespignani, A.: Activity driven modeling of time varying networks. Sci. Rep. **2** (2012) 469
5. Zino, L., Rizzo, A., Porfiri, M.: Continuous-time discrete-distribution theory for activity-driven networks. Phys. Rev. Lett. **117** (2016) 228302
6. Abaid, N., Porfiri, M.: Consensus over numerosity-constrained random networks. IEEE Trans. Autom. Control **56**(3) (2011) 649–654
7. Buscarino, A., Fortuna, L., Frasca, M., Gambuzza, L., Nunnari, G.: Synchronization of chaotic systems with activity-driven time-varying interactions. J. Complex Netw. **6**(2) (2018) 173–186
8. Ogura, M., Tagawa, J., Masuda, N.: Distributed agreement on activity driven networks. In: Proc. Am. Control Conf. (2018) 4147–4152
9. Zino, L., Rizzo, A., Porfiri, M.: Consensus over Activity Driven Networks. (2019) *In Review*.
10. Hasanyan, J., Burbano Lombana D.A., Rizzo, A., Porfiri, M.: Leader-follower consensus on activity-driven networks. (2019) *In Review*.

Restructuring mechanisms of the hierarchical networks between PubMed MeSH terms

Gergely Palla[1], Péter Pollner[1], Dániel Zagyva[2], and Sámuel G. Balogh[2]

[1] MTA-ELTE Statistical and Biological Physics Research Group,
Pázmány P. stny. 1/A, 1117 Budapest, Hungary,
pallag@hal.elte.hu, pollner@hal.elte.hu
[2] Dept. of Biological Physics, Eötvös University,
Pázmány P. stny. 1/A, 1117 Budapest, Hungary
zagyva@hal.elte.hu, balogh@hal.elte.hu

1 Introduction

Signs of hierarchical organisation can be often observed in complex networks, supported by various studies with subjects ranging from flocks of various species [1] through social interactions [2] to scientific journals [3] and on-line news content [4]. In most of the cases, real networks are constantly evolving in time, and some relevant aspects of the laws forming the structure of these systems have already been uncovered in the scientific literature. One of the most well known example is the preferential attachment rule for growing scale-free networks, corresponding to the key concept of the Barabasi-Albert model [5], which was also detected by empirical studies of network data [6, 7]. In a very recent work, along a similar line, we have examined the statistical features of the restructuring mechanisms in networks with a hierarchical structure [8], where the main goal was to detect preference or anti-preference during the different attachment and detachment events over the time evolution.

The networks we studied correspond to the hierarchies between the Medical Subject Headings (MeSH terms) provided by the NCBI to help searching in the PubMed publication database (comprising more than 29 million citations for biomedical literature) at various levels of specificity. The MeSH terms are sorted into 16 hierarchies (labelled A, B, C, etc.), and at the top of the hierarchies we find very broad headings such as "Organisms" or "Information Science", whereas more specific headings are found at deeper levels. Due to the rapidly developing nature of the medical-, biochemical- and biological sciences, the set of available MeSH terms are yearly updated by the curators of PubMed.

2 Methods

In order to briefly describe our method for detecting preference with regard to some node property x, let us consider first only two consecutive time steps. We denote the probability distribution of x at the initial state by $p(x)$, and the complementary cumulative distribution of x as $Q(x) = \sum_{x' \geq x} p(x')$. By taking the ratio between $w(x)$, corresponding to the number of chosen nodes by the considered attachment procedure for which

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the property value is at least as large as x and $Q(x)$ resulting in $W(x) = w(x)/Q(x)$, we obtain a function that is constant if the attachment is uniform in x, since in this case $w(x)$ and $Q(x)$ are simply proportional to each other for any x. In contrast, if larger values of x are preferred, the shape of $W(x)$ becomes increasing as a function of x, whereas in the opposite case, when the attachment/detachment prefers lower values of x, the shape of $W(x)$ becomes decreasing. Due to its simple construct, the expected value and variance of $W(x)$ under uniform random choice (where the attachment is independent of x) can be calculated analytically, for details see Ref.[8].

To measure the preference of the attachment procedure over the whole period of time steps in the empirical data, for every time step t (except for the last) we can measure the complementary cumulative distribution $Q_t(x)$, and compare it to $w_t(x)$, denoting the number of nodes having a property value at least as large as x selected by the given attachment mechanism between t and $t+1$. By aggregating their ratio, we can define

$$W_{\mathrm{emp}}(x) = \sum_{t=1}^{t_{\max}-1} \frac{w_t(x)}{Q_t(x)}. \tag{1}$$

The obtained curve can be then compared to the expected value of the random variable corresponding to the sum of the supposed $W(x)$ under the assumption of independence from x, which we can denote by $W_{\mathrm{rand}}(x)$.

3 Results

We applied the method outlined in the previous section to study the time evolution of the MeSH hierarchies with a system size exceeding 1000 nodes during the whole recorded time period, focusing on the following properties: number of children (out degree), number of parents (in degree), total number of descendants, total number of ancestors. What makes the problem non-trivial is the rather high number of different possible attachment and detachment event types that can occur during the time evolution. In terms of the changing links we have two large categories: added (new) links and deleted links. When examining the endpoints of added links, both the source and the target can be either an already existing (old) node, or a new node, thus, there are altogether 4 types of added links. The case of deleted links is much simpler in this respect, as both endpoints must correspond to old nodes. Therefore, there are in total 5 different possibilities for changes in the connections. However, when examining the possible effect of a given node property on the likelihood that the node is going to take part in an attachment/detachment event, we also have to specify whether the node is the source or the target of the involved link. Thus, for any node property of interest we can examine 10 different scenarios over the time evolution of the hierarchies.

As an illustration of the obtained results, in Fig.1 we show the measured $W_{\mathrm{emp}}(x)$ and corresponding $W_{\mathrm{rand}}(x)$ curves for two cases. According to Fig.1a, the attachment of new links pointing from old nodes to new comers shows a strong preference with respect to the total number of descendants of the source node in case of hierarchies D and C. In contrast, Fig.1b indicates that the attachment of new links appearing between old nodes shows anti-preference with respect to the number of ancestors of the source node.

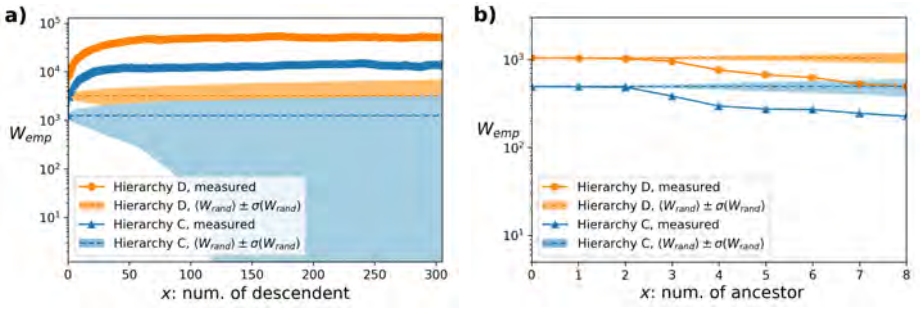


Fig. 1. Measuring preference under restructuring events. In both panels we compare $W_{\mathrm{emp}}(x)$ defined in (1) to the mean and standard deviation of $W(x)$ for random events, indicated by dashed lines in shaded areas. a) Results for the total number of descendants of source nodes in attachments of new links pointing from old nodes to new nodes in hierarchies D (orange) and C (blue). b) $W_{\mathrm{emp}}(x)$ for the number of ancestors of source nodes on new links appearing between old nodes for the same hierarchies as in panel a).

The results for the further attachment types and the other hierarchies are given in Ref.[8]. Based on those, we could observe strong signs of preference with respect to the number of children of the source node for both the addition of new links pointing from old nodes to new ones, and for the deletion of already existing links between old nodes. In parallel, we saw anti-preference with respect to the number of ancestors of the source node for all possible link change types. Interestingly, if the node acts as the target of the changing link, we could observe both preference and anti-preference with respect to the number of ancestors for the different link change types [8]. In conclusion, our results indicate that time evolution of these systems is far more complex compared to simple preferential attachment models, providing very interesting future challenges for modelling and further statistical analysis.

References

1. Nagy M, Ákos Z, Biro D, Vicsek T. Hierarchical group dynamics in pigeon flocks. Nature. 2010;464:890–893.
2. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. Phys Rev E. 2003;68:065103.
3. Palla G, Tibély G, Mones E, Pollner P, Vicsek T. Hierarchical networks of scientific journals. Palgrave Communications. 2015;1:15016.
4. Tibély G, Sousa-Rodrigues D, Pollner P, Palla G. Comparing the Hierarchy of Keywords in On-Line News Portals. PLoS ONE. 2016;11:e0165728.
5. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999;286:509–512.
6. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A. 2002;311:590–614.
7. Pollner P, Palla G, Vicsek T. Preferential attachment of communities: The same principle, but a higher level. Europhys Lett. 2006;73:478–484.
8. Balogh SG, Zagyva D, Pollner P, Palla G. Time evolution of the hierarchical networks between PubMed MeSH terms. PLoS ONE. 2019;14:e0220648.

Community detection in non-stationary temporal networks

Alexandre Bovet[1,2] and Jean-Charles Delvenne[1], Renaud Lambiotte[2]

[1] ICTEAM, Université catholique de Louvain, Avenue George Lemaître 4, 1348 Louvain-la-Neuve, Belgium,
alexandre.bovet@uclouvain.be,
WWW home page: alexbovet.github.io
[2] Mathematical Institute, University of Oxford, Oxford, UK

1 Introduction

Many temporal networks exhibit non-stationary dynamics, such as cyclical patterns due to daily, weekly, seasonal or yearly cycles, increase or decrease in population size or drastic change of dynamical regime. Several works have generalized existing community detection methods for static networks to temporal networks (e.g. [1–5]), but they usually rely on the assumption of an underlying stationary process, or sequences of different stationary epochs, and a null model corresponding to the stationary state of the process. Here, we propose a first-principle method allowing to take into account continuous time temporal networks, interactions that may have a duration and systems that non-necessarily reach a steady state, or follow a sequence of stationary states.

2 Results

Our approach is based on the concept of the stability of a network partition [6, 7] generalized to temporal networks with non-Markovian and non-stationary dynamics.

Given a temporal network with a fixed number of nodes N and a set of directed edges $e = (v_s, v_t, t_s, \Delta t)$ where v_s and v_t are the source and target vertices, respectively, t_s is the time at which the edge becomes active and Δt is the duration of the edge, we compute the matrix of transition probabilities with element $T_{ij}(t_1, t_2)$ equal to the probability of going from node i at t_1 to node j at t_2 by considering a continuous time random walk with rate λ that is constrained by the activation of the edges. Communities are then defined as groups of nodes that retain the flow of walkers the most over a given time span (t_1 to t_2). They are found by optimizing the quality function that we call the *flow stability*:

$$r^{\text{flow}}(t_1, t_2; \mathbf{H}) = \text{trace}\left[\mathbf{H}^T \mathbf{S}(t_1, t_2)\mathbf{H}\right], \tag{1}$$

where, $\mathbf{S}(t_1, t_2) = \text{diag}(\mathbf{p}(t_1))\mathbf{T}(t_1, t_2) - \mathbf{p}(t_1)^T\mathbf{p}(t_2)$ is the autocovariance matrix of the process, $\mathbf{p}(t)$ is the probability density vector of the random walk at time t and $\mathbf{H}$ is an indicator matrix that encodes which node belong to which community. The optimization can be performed, for example, with the Louvain algorithm [8]. The rate

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

of the random walk, λ, plays the role of a resolution parameter, allowing to detect communities at all scales. Interestingly, in the case of static undirected networks, this expression evaluated at stationarity reduces to the static Markov Stability [6, 9] which is equal to the classic Newman-Girvan Modularity [10] for a Markov time (resolution parameter) equal to one. In the case of directed static network, considering one step of a discrete-time random walk, the flow stability reduces to a standard generalization of modularity to directed network ($Q^{\mathrm{d.}} = \frac{1}{m}\sum_{ij}\left(A_{ij} - (k_i^{\mathrm{out}}k_j^{\mathrm{in}})/m\right)\delta\left(c_i, c_j\right)$).

We show how the autocovariance matrix is asymmetric in general, whether the edges of the temporal networks have a direction or not. Indeed, the time ordering of events can result in different probabilities for going from a particular node i at t_1 to a node j at t_2 than going from j at t_1 to i at t_2 [11], even if each event allows walkers to travel in both directions. To capture this asymmetry, we propose to describe the communities in temporal networks with two partitions: the *source* and *target* partitions. clustering the rows and columns of the autocovariance matrix separately (see Fig. 1).

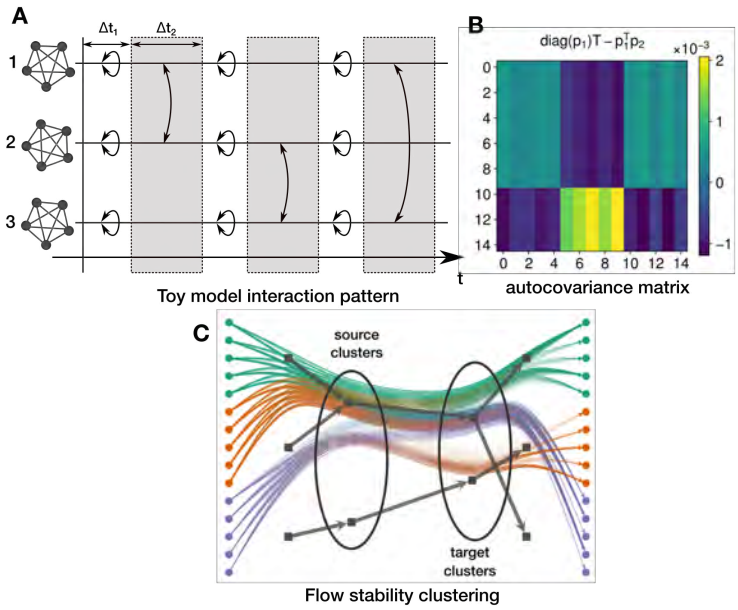


Fig. 1. Temporal flow clustering. (**A**) We consider a toy model made of three groups of 5 nodes. Nodes activations follow a Poisson process and edges durations are drawn from an exponential distribution. The system follows two types of successive interactions: $I1$) during Δt_1 each vertex interacts with other vertices of its own group with the largest probability; $I2$) during Δt_2 the vertices of two of the groups interact with one another with the largest probability. (**B**) The autocovariance matrix we derive allows to put into evidence the temporal communities structure and reveals the asymmetry of the system arising from the specific time ordering of events. (**C**) Clustering found by our approach showing how the time-asymmetric flow of walkers is clustered in source communities and target communities.

Summary. Our method generalizes the concept of modularity [10] of a network partition for general temporal networks [5], over a given temporal interval, by taking into account time respecting paths, capturing the asymmetry created by the time ordering of events and allowing to consider multiple scales of the system. We consider applications of our method to a toy model and several real-world examples, such as an extensive contact network of free-living wild mice [12].

References

1. Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. & Onnela, J.-P. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* **328**, 876–878 (2010). URL http://www.sciencemag.org/cgi/doi/10.1126/science.1184819.
2. Petri, G. & Expert, P. Temporal stability of network partitions. *Physical Review E* **90**, 022813 (2014). URL https://link.aps.org/doi/10.1103/PhysRevE.90.022813.
3. Aslak, U., Rosvall, M. & Lehmann, S. Constrained information flows in temporal networks reveal intermittent communities. *Physical Review E* **97**, 062312 (2018). URL https://link.aps.org/doi/10.1103/PhysRevE.97.062312. 1711.07649.
4. Peixoto, T. P. & Rosvall, M. Modelling sequences and temporal networks with dynamic community structures. *Nature Communications* **8**, 582 (2017). URL http://www.nature.com/articles/s41467-017-00148-9.
5. Rossetti, G. & Cazabet, R. Community Discovery in Dynamic Networks. *ACM Computing Surveys* **51**, 1–37 (2018). URL http://dl.acm.org/citation.cfm?doid=3186333.3172867.
6. Delvenne, J. C., Yaliraki, S. N. & Barahona, M. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences* **107**, 12755–12760 (2010). URL http://www.pnas.org/cgi/doi/10.1073/pnas.0903215107.
7. Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N. & Barahona, M. Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit. *PLoS ONE* **7**, e32210 (2012). URL https://dx.plos.org/10.1371/journal.pone.0032210.
8. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008). URL http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52.
9. Lambiotte, R., Delvenne, J.-C. & Barahona, M. Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. *IEEE Transactions on Network Science and Engineering* **1**, 76–90 (2014). URL http://ieeexplore.ieee.org/document/7010026/.
10. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Physical Review E* **69**, 026113 (2004). URL https://link.aps.org/doi/10.1103/PhysRevE.69.026113.
11. Scholtes, I. *et al.* Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications* **5**, 5024 (2014). URL http://www.nature.com/articles/ncomms6024.
12. König, B. *et al.* A system for automatic recording of social behavior in a free-living wild house mouse population. *Animal Biotelemetry* **3**, 39 (2015). URL http://www.animalbiotelemetry.com/content/3/1/39.

Constant State of Change: Engagement Inequality in Temporal Dynamic Networks (Extended Abstract)

Hadar Miller and Osnat Mokryn

University of Haifa, Israel

1 Introduction

Temporal measures of engagement are of interest as they give a measure of member participation, interest, influence, dominance, and more [1]. In organizations, where frequent changes were found to be the norm [2], following the temporal intensity and dominance of the interactions can help in identifying fluctuations in involvement and engagement prior, during, and after a planned organizational change, as well as assess the reactions to a shock. These temporal measures are of interest also in the case of online social networks engagement, where participation was found to be dominated by a few. Yet, participants change their active role in the network and their engagement over time [3]. Currently, it is unclear whether these changes affect the temporal measures of network activity. Hence, we set to understand the change in the average intensity of interactions and the variance in them. The distribution of the intensity of interactions, also referred to as ties' strength, has long been recognized as a fundamental property [4, 5]. We continue to define indices of average connection intensity and nodal dominance inequality in temporal networks. A measure of average intensity of the edge interactions in a network differs from average nodes' strength, as the measure should not favor the number of active connections a node has.

Temporal Intensity Level index: Centrality measure in weighted networks is defined in [5] as follows: $C_D^{w\alpha}(i) = k_i^{(1-\alpha)} \cdot s_i^{\alpha}$, where $\alpha \in [0,1]$ is the tuning parameter, k_i is the number of nodes the focal node i is connected to, and s_i is its weighted degree. s_i is computed by: $s_i = \sum_i^N w_{ij}$, where N is the total number of nodes in this network, and w_{ij} is a non-zero value for the strength of edges that disseminate from the focal node i. Taking a network-wide approach, we define the weighted sum as follows:

$$\phi_\alpha = \sum_{i=1}^{N} C_D^{w\alpha}(i) = \sum_{i=1}^{N} k_i^{(1-\alpha)} \cdot s_i^{\alpha} \tag{1}$$

The metric $\phi_{\alpha=0}$ corresponds to the number of edges in the graph; Alternatively, the metric $\phi_{\alpha=1}$ corresponds to the sum of all edge weights in the network, that is, the overall intensity of interactions in a network. The Temporal Network Intensity index for networks is the ratio between the overall intensity of edge interactions in the network and the binary number of edges, over a predefined window of time[1]:

$$\psi(G_\tau) = \frac{\phi_{\alpha=1}(G_\tau)}{\phi_{\alpha=0}(G_\tau)} \tag{2}$$

Where $G_\tau, \tau \in [1..T]$ is a sequence of graphs representing consecutive network snapshots in a period T. $\psi \geq 1$ holds for all graphs.

[1] A discussion on the length of the time window is outside the scope of this abstract.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Temporal Dominance Inequality: In organizations, when a change is introduced, high interactions can be found among its supporters and opposers, but there might be a silent majority. Understanding the level of inequality in the intensity of the participation can aid in understanding the balance between change-involved members versus those who are not [2]. We measure the inequality in nodal interactions dominance utilizing the Gini inequality index [6] for measuring income inequality.

2 Results

We gathered the temporal interactions from six real world networks [2]. For each of the datasets we calculate the weekly temporal network intensity, as defined in Equation 2. The results, as appear in Figure 1 are surprising. All networks exhibit a rather stable temporal behavior in their intensity, regardless of the fluctuations in size. It is also interesting to note that although the Intensity is not bounded in value, in all these networks the average intensity is low. For example, the Facebook network, on the lower left panel, show a steady increase in network size from several hundreds up to more than 10000 weekly participants. (minimal intensity is calculated from zero as explained above). We get similar results when measuring the temporal dominance (Gini index) in these networks. The measured values are in the range of $[0.4, 0.7]$ for all datasets. Intuitively, an Erdös-Rényi (ER) random network would yield very low inequality values, as all nodes have a similar chance for communicating, and a pure Preferential Attachment (PA) network would give a very high inequality value. Figure 2 denotes the cumulative

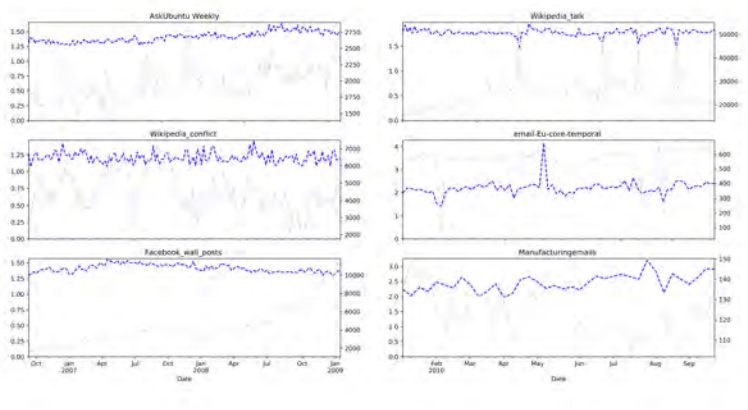
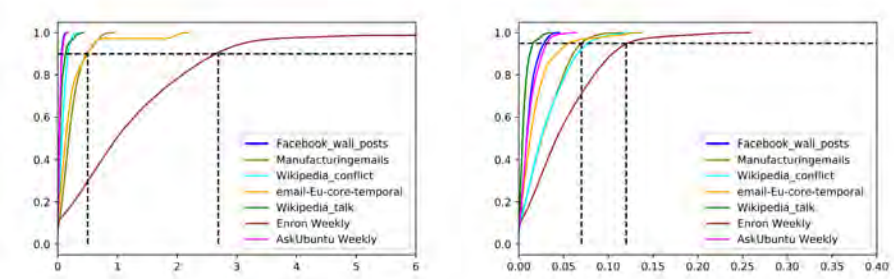


Fig. 1. Temporal average intensity for the six datasets, denoted by the blue line with the values on the left y-axis. The light grey dashed line corresponds to the temporal size of the network, denoted by the right y-axis.

distribution of the relative change in the measured indices between every two consecutive weeks for each dataset. In all networks but Enron more than 80% of the changes are

[2]The datasets are: AskUbuntu forum (198 weeks); Facebook Wall Posts (124 weeks); Wikipedia Conflict (156 weeks); Wikipedia Talk (132 weeks); Manufacturing Emails (38 weeks); EU Research Institutional Emails (74 weeks).

of less than 15%. the Enron network, used often for change point detection, is different from the other networks examined in terms of the range of Temporal Network Intensity index and the percentage of changes measured in the index. The network displays Temporal Network Intensity in the range of $3.0 - 12.0$, well above the index range for the other networks. In addition, the index volatility is very high and the changes between weekly measurements are high. The Temporal Dominance Inequality, as presented in Figure 2(B), while is similar in range to that of other networks, also shows high volatility compared to the other networks. Our results determine that networks differ by the



(a) CDF of weekly change in average intensity (b) CDF of weekly change in Dominance

Fig. 2. The cumulative distribution of the weekly relative change for each dataset in the measured. ((a)) Temporal Network Intensity and ((b)) Temporal Dominance Inequality.

engagement indices we defined. To further verify this result, we ran a classification experiment over the weekly indices, and find that the classifier can classify the indices tuples to their corresponding network with high validity.

Summary. Our surprising results are that for most emails and forum networks checked, the indices introduced were *stationary*, implying a steady state. The robustness of the indices regardless of significant size changes of the underlying network in time, is intriguing. For example, when the size of the network decreases, in a process of preferential detachment it is expected that the level of engagement and hence the indices would be also effected. Lastly, our result show that the indices we devised fluctuated significantly in a network that was dealing with a shaky situation that let to the company's disintegration.

References

1. A. Li, S.P. Cornelius, Y.Y. Liu, L. Wang, A.L. Barabási, Science **358**(6366), 1042 (2017)
2. W.W. Burke, *Organization Change: Theory and Practice* (SAGE Publications, 2017)
3. J. Nielsen, http://www. useit. com/alertbox/participation_inequality. html (2006)
4. A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, Proceedings of the national academy of sciences **101**(11), 3747 (2004)
5. T. Opsahl, F. Agneessens, J. Skvoretz, Social networks **32**(3), 245 (2010)
6. C. Gini, The Economic Journal **31**(121), 124 (1921)

Uncertainty in the critical threshold for dynamics on complex networks

Lluís Arola-Fernández[1], Guillem Mosquera-Doñate[2,3], Benjamin Steinegger[1],, and Alex Arenas[1]

[1] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007, Tarragona Spain
alexandre.arenas@urv.cat,
[2] Center for Complexity Science, University of Warwick, Coventry, UK.
[3] The Alan Turing Institute, London, UK.

1 Introduction

The study of dynamical processes running on top of complex networks has become a key problem in many research fields, ranging from the microscopic realm of genes and neurones to the large realm of technological and social systems [1]. However, in many practical situations, there is a lack of precision in the measurements and also intrinsic fluctuations may be present in the interactions of the network. These sources of uncertainty in the structure affect dramatically the dynamical properties (as the critical threshold of a macroscopic phase transition or the stability of the dynamical attractor), and they should be taken into account when making analytical predictions from the available data.

Following this line, we study the uncertainty in the critical threshold of a general dynamical process on top of a complex network, when it is induced by microscopic noise in the intensity of the connections among the units. Here, we present an analytical formalism that captures the main statistics of the threshold when affected by white gaussian noise in the weights of the network. Our theory has a very good agreement against simulations and the results show how the underlying structure of interactions plays a central role in the way the microscopic noise is propagated through the macroscopic threshold. In particular, the theory predicts the existence of optimal structures that are able to amplify significantly the critical range only due to small fluctuations in the weights.

2 Results

We consider a network with a fixed structure of links that capture the presence of connections among units and we let the intensity of the links (the weights) to be affected by random fluctuations. For simplicity, we assume that the noise is gaussian and uncorrelated (white noise) where each weight is drawn from a normal distribution $N(\mu, \sigma)$. The main goal is to understand how this microscopic noise affects the value of the critical

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

point in a dynamical process running on top of the network.

For a variety of dynamical processes running on top of complex networks (including synchronization, spreading dynamics and spin models) the critical threshold K_c is estimated in terms of the inverse of the largest eigenvalue λ_{max} of the adjacency matrix $\mathbf{A}$ [2–4]. In order to study the exact statistics of K_c in the presence of noise with 0 mean, one should use the tools from Random Matrix Theory [5]. However, for sparse networks with arbitrary degree distributions, it becomes very challenging to obtain analytical results in this context. Also, since we are particularly interested in the scenario where the mean of the interactions is not zero $(\mu > 0)$, an alternative approach is required.

We tackle this problem by applying an error propagation method to the mean-field approximation of the threshold [6]. This method, although being approximate, gives surprisingly accurate results and provides closed form expressions that facilitate our understanding on the problem. We are able to derive closed-form expressions for the mean and the variance of the critical threshold depending on the noise parameters and the moment of the degree distribution of the underlying network [7]. In Fig.(1), we show the accuracy of the mean-field approximation in capturing the distribution of the critical threshold (left) and the performance of the theoretical expressions for a fixed Erdös-Rényi network (right). The results are also tested in many empirical networks showing good agreement against simulations (not shown in the abstract).

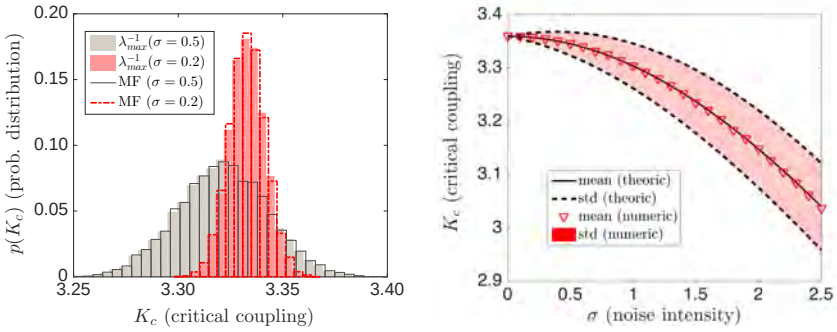


Fig. 1. Left: Empirical (areas) and MF (lines) histograms for the distribution of the K_c in a fixed Erdös-Rényi network with $N = 200$, $p = 0.3$, $K_0 = 1$, $\mu = 1$ for two different noise intensities σ with 1000 realizations. The statistics are indeed affected by the noise and the MF approx. accurately estimates the whole distribution of K_c. Right: Numeric vs theory: mean and standard deviation of the critical threshold depending on the noise intensity σ for a fixed Erdös-Rényi network with $N = 200$, $p = 0.3$, $\mu = 1$ and 1000 independent realizations.

Furthermore, our theoretical results show that the fluctuations in the critical point depend non-linearly on the moments of the degree distribution and the noise parameters. We were able to find which are the structures that maximize or minimize the critical fluctuations, for a given amount of noise. This result finds implications also in the context of adaptation and evolution of many biological systems [8]: some structures

are able to increase their critical range (and therefore the variety of macrostates) only by small fluctuations of the weights, without altering the underlying structure of links.

We propose an error propagation method to analytically quantify the macroscopic uncertainty on the critical threshold of a dynamical process when induced by white noise on the coupling weights of a network. The method is tested with good accuracy in synthetic and empirical data. The results unveil several interesting noise-amplifying properties of the networks and the method can be used in practical situations, to quantify the error made by theoretical predictions due to uncertain measurements.

References

1. M. Newman, Networks: An Introduction (Oxford University Press, Inc., New York, NY, USA, 2010).
2. S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Rev. Mod. Phys. 80, 1275 (2008).
3. A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, Phys. Rep. 469, 93 (2008).
4. R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. 87, 925 (2015).
5. S. Jalan and J. N. Bandyopadhyay, Phys. Rev. E 76, 046107 (2007).
6. J. G. Restrepo, E. Ott, and B. R. Hunt, Phys. Rev. E 76, 056119 (2007).
7. L. Arola-Fernández et al. (under preparation).
8. D. B. Larremore, W. L. Shew, and J. G. Restrepo, Phys. Rev. Lett. 106, 058101 (2011).

Effective Dynamics on Complex Networks

Flvio L. Pinheiro[1,4], Jorge M. Pacheco[2,4], and Francisco C. Santos[3,4]

[1] Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal,
[2] INESC-ID and Instituto Superior Tcnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal
[3] CBMA and Departamento de Matemtica e Aplicaes, Universidade do Minho, 4710-057 Braga, Portugal
[4] ATP-group, P-2744-016 Porto Salvo, Portugal

Complex networks are known to profoundly affect the processes they support. Some of the most complex processes investigated to date on networks are related with behavioural dynamics and decision-making. These are often abstracted by means of social dilemmas of cooperation, such as the *Prisoners Dilemma* (PD). In that context, despite the higher returns of mutual cooperation, rational agents are paradoxically expected to mutually defect, thus the dilemma. An evolutionary population dynamics approach to game theory, where agents revise their behaviour based on the perceived success of others, provides one of the most sophisticated examples of complex dynamics in which the role of the underlying network topology proves key to determine the evolutionary outcome of a population. For instance, when cooperation is modeled as a PD, cooperation may emerge (or not) depending on how the population structure [9].

However, the precise link between the local self-regarding actions and the population-wide dynamics that might lead to a collective cooperative scenario on structured populations has been hard to establish. Indeed, past studies have mostly focused on the analysis of the evolutionary outcome of cooperation – either by means of the numerical analysis of steady states or by the analytical determination of the conditions that lead to fixation – thus lacking a characterization of the self-organization process by which one of the strategies out competes the other. Here we report on a numerical approach [5, 9, 10, 4, 7] that unveils the link between individual and collective behavior in evolutionary games on structured populations.

To that end we define a time-dependent variable – the *Average Gradient of Selection* (AGoS) – and use it to track the self-organization of cooperators when co-evolving with defectors. In finite well-mixed populations the gradient of selection, $G(k) = T^+(k) - T^-(k)$, can be computed analytically as the difference between the probabilities of increasing ($T^+(k)$) and decreasing ($T^-(k)$) the number of cooperators by one, for a population with k cooperators. It is impossible to compute $G(k)$ analytically for arbitrary network structures [2], in that sense, the AGoS provides a numerical account of the same variables, offering the change in time of the frequency of cooperative traits under selection. The AGoS can be computed for arbitrary intensity of selection, arbitrary population structure, and arbitrary game parameterization.

Overall, we show how behavioral dynamics of individuals facing a cooperation dilemma in structured populations can be understood as though individuals face a different dilemma in a well-mixed (*i.e.*structured-less) population. As illustrated in Fig. 1, homogeneous networks promote a coexistence dynamics between cooperators and de-

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

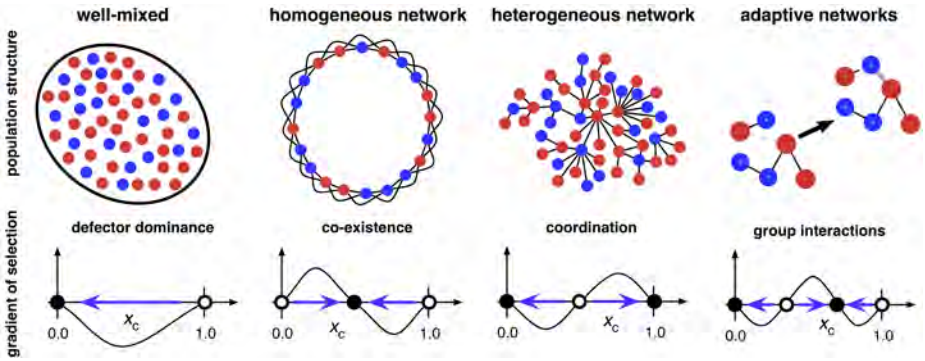


Fig. 1. The Average Gradient of Selection (AGoS) provides a characterization of the change in time of the fraction of cooperators under natural selection, being positive (negative) when the fraction of cooperators tends to increase (decrease). While in well-mixed populations, the tragedy of the commons ($x_C = 0$) emerges as the only stable fixed point, homogeneous networks favor the co-existence of cooperators and defectors, whereas degree heterogeneous networks creates two basins of attraction, as if agents would be locally facing a coordination dilemma. Adaptive network structures lead to the emergence of a two interior fixed points, a dynamical fingerprint of N-Person games that involve group social dilemmas.

fectors – akin to a Snowdrift game – whereas strongly heterogeneous networks prompt a coordination between them, similar to the Stag-hunt game. In other words, while agents locally perceive and play a PD, globally the dynamics of the population resembles the one obtained from a completely different game, as if, individuals would be locally facing a different dilemma.

In [6] use the AGoS to show that contrary to what happens in heterogeneous populations that generate a coordination dynamics for a broad range of selection pressure values, on homogeneous networks the population-wide dynamics depends on the intensity of selection: under strong selection they favour a co-existence like dynamics while under weak selection we recover the well-mixed scenario of a PD-like dynamics which leads to the demise of cooperation (Fig. 1). [4] have shown the existence of an optimal range of network heterogeneity that optimizes the evolutionary cooperative outcome of a population, reinforcing the idea of the sensitivity of evolutionary games to the underlying features of population structure. Moreover, we were able to identify the existence (on several types of networks) of an optimum level of selection pressure for which cooperation is maximised. The underlying process that leads to this result differs from homogeneous to heterogenous networks. In the first class of networks the optimal selection pressure is associated with the ability of cooperators to form and sustain clusters, while on the second class it is the result of a decoupling in the distribution of intensities of selection between pairs of agents that is present from the natural diversity of fitness values [10] in the population.

When the co-evolution of both strategies and network structure is considered, the range of social dilemmas where cooperation can thrive expands. In [7] we show that, when individuals engage locally in PD games, we observe that adaptive networks give rise to the emergence of population-wide dynamics that is akin to what we find in lo-

cal games that involve group interactions (N-Person Games) with non-linear returns [3]. Interestingly, such results means that adaptive social structures entwine individuals decisions in scenarios that extend their dyadic relationships.

Underlying these emergent phenomena are the natural build up of peer-influenced correlations between individuals behaviors along nodes of the network. We have shown that such correlations emerge from the pairwise learning dynamics in populations with network mediated interactions [8]. These patterns are characterized by positive correlations among the strategies of individuals up to two or three links of separation. Our results nicely match and extend our understanding of previous empirical studies that found similar peer-influence patterns in social networks [1].

The application of the AGoS is not limited to 2-person games. In fact, as discussed in [10], heterogeneous network structures create multiple internal equilibria when individuals face public goods dilemmas, departing significantly from the reference scenario of a well-mixed populations. Finally, we would like to stress that the scope and importance of this methodology goes beyond the present application to evolutionary games on graphs. The principles can be used to extract any dynamical quantity that describes a process (as long as it is a Markov process) taking place on a network such as the outbreak of epidemics or the opinion diffusion.

References

1. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. New England Journal of Medicine 357(4), 370–379 (2007)
2. Ibsen-Jensen, R., Chatterjee, K., Nowak, M.A.: Computational complexity of ecological and evolutionary spatial dynamics. Proc Natl Acad Sci USA 112(51), 15636–15641 (2015)
3. Pacheco, J.M., Santos, F.C., Souza, M.O., Skyrms, B.: Evolutionary dynamics of collective action in n-person stag hunt dilemmas. Proc Royal Soc B: Biol. Sciences 276(1655), 315–321 (2008)
4. Pinheiro, F.L., Hartmann, D.: Intermediate levels of network heterogeneity provide the best evolutionary outcomes. Sci Rep 7(1), 15242 (2017)
5. Pinheiro, F.L., Pacheco, J.M., Santos, F.C.: From local to global dilemmas in social networks. PloS ONE 7(2), e32114 (2012)
6. Pinheiro, F.L., Santos, F.C., Pacheco, J.M.: How selection pressure changes the nature of social dilemmas in structured populations. New J Phys 14(7), 073035 (2012)
7. Pinheiro, F.L., Santos, F.C., Pacheco, J.M.: Linking individual and collective behavior in adaptive social networks. Phys Rev Lett 116(12), 128702 (2016)
8. Pinheiro, F.L., Santos, M.D., Santos, F.C., Pacheco, J.M.: Origin of peer influence in social networks. Phys Rev Lett 112(9), 098702 (2014)
9. Santos, F.C., Pinheiro, F.L., Lenaerts, T., Pacheco, J.M.: The role of diversity in the evolution of cooperation. J Theor Biol 299, 88–96 (2012)
10. Santos, M.D., Pinheiro, F.L., Santos, F.C., Pacheco, J.M.: Dynamics of n-person snowdrift games in structured populations. J Theor Biol 315, 81–86 (2012)

A Minimal Co-evolving Voter Model on Simplicial Complexes

Leonhard Horstmeyer[1,3] and Christian Kuehn[1,2]

[1] Complexity Science Hub Vienna, Josefstädterstrasse 39, A-1090 Vienna, Austria.
horstmeyer@csh.ac.at
[2] Faculty of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching München, Germany. ckuehn@ma.tum.de
[3] Basic Research Community for Physics, Mariannenstr. 89,04315 Leipzig, Germany.

1 Introduction

Network theory has played a crucial role in enhancing our understanding of polarization, segregation, fragmentation, hierarchical stratification, and other phenomena related to opinion formation and propagation[1]. The underlying paradigm is the formal represention of social systems by networks in which the nodes correspond to agents and the links to binary relations amongst them. Recently, higher-order relations have started to appear as a new focus in the analysis of complex network data sets [2]. Such inherently social phenomena as peer groups, contracts, institutionalisation and state formation demand for an inclusion of these higher-order interactions into models and theoretical descriptions. One may even argue that their omission fails to capture the essence of social systems in the same way the linear models fail to capture the essence of natural processes.

Here we consider one of the classical models of opinion formation which exhibits fragmentation, the co-evolving voter model[3], and propose an extension to higher-order interactions[4]. We recall the classical co-evolving voter model in which agents are endowed with one of two possible opinions, say +1 or -1. They are connected via links, forming a network. Both the opinion states and the network itself evolve, accounting for an adaptation and thus giving rise to the co-evolutionary nature of the model. Two connected nodes with opposing opinions either homogenize their opinion with probability $1-p$ or they rewire their connection with probability p. The persuading or rewiring node is chosen at random. One observes that either one of the opinions wins in the long run or the network fragments into two disjoint communities of opposing opinions. There is a critical rewiring probability p_c above which the network fragments.

We propose to model peer groups by simplices and extend the classical co-evolving voter model by a majority rule that models peer pressure [4]. Here we describe a minimal version in which peer groups are 2-simplices, but extensions to n-simplices are straightforward. A 2-simplex is a triadic relation of three nodes that requires binary relations amongst each of its vertices. One may think of it as a filled out triangle. Typically in a peer group all members are also friends with each other, justifying our modeling assumption. A system consisting of nodes, edges and 2-simplices is called a simplicial 2-complex.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

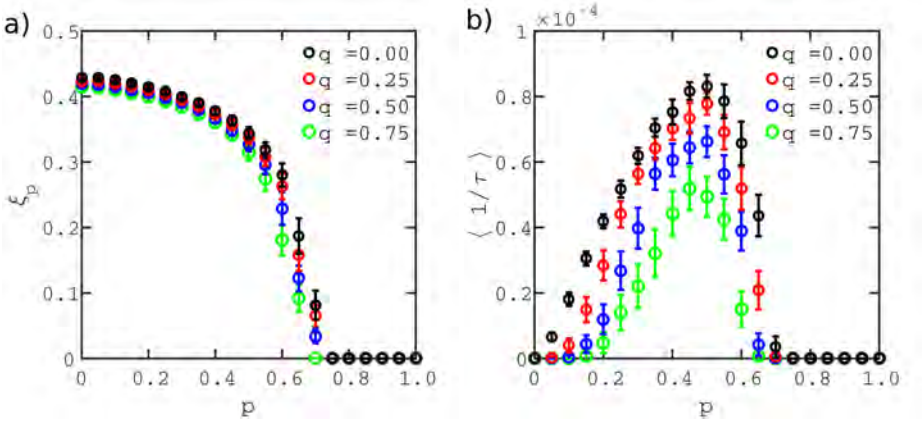


Fig. 1. We show a) the order parameter ξ_p of the co-evolving voter model on simplicial complexes and b) the average inverse depletion time of triangles $\langle 1/\tau \rangle$, respectively for various rewiring probabilities and peer pressures. The simplicial complex are randomly generated for $N = 500$ nodes, a mean degree $\mu = 8$ and 2-simplex-per-edge degree $s = 0.2$.

The majority rule states that the majority opinion convinces the minority opinion when an active edge inside a simplex is chosen. The majority rule is applied with probability q, thus $q = 0$ corresponds to the classical co-evolving case. In summary the model is described by the following update rule: At each time step an edge e is chosen. If this edge connects the same opinions nothing happens. If it connects opposing opinions, either the classical rules apply with probability $1 - q$ or the majority rule applies with probability q. If the majority rule applies, then one of the simplices attached to that edge is chosen for the persuasion and if none is present, i.e. in the absense of a peer group, the classical update rule applies. Whenever a simplex is destroyed by a rewiring event a randomly chosen triangle is converted into a simplex.

This minimal extension allows us to study the effect of peer pressure in voter processes. It also serves the purpose of studying evolving simplical complexes as such by means of a simple model.

2 Results

We conduct numerical simulations supported by calculations. First we find that higher peer pressures accelerate the fragmentation process and the fragmentation itself already occurs at lower rewiring probabilities. In Figure 1a) we show the order parameter ξ_p for various peer pressures $q \in \{0, 1/4, 2/4, 3/4\}$, where ξ_p is the maximal quasi-stationary density of inhomogeneous (sometimes called active) links [3]. The simulations are initialized by random simplicial complexes with $N = 500$ nodes, a mean degree of $\mu = 8$ and a low simplex-per-edge density of $s = 0.2$. Simplices are distributed uniformly at random over the set of vertex-triplets. Despite the low simplex density one may see clearly the effect of an earlier fragmentation for higher peer pressures. Apart from the early fragmentation transition one may look at the depletion rate of active edges and

also – for values of p below p_c – at the drift velocity towards one of the single-opinion states. In both cases we find that the peer pressure increases the respective velocities.

Secondly, we find that there is a multiscale hierarchy of time scales that correspond to the order of the simplex. The evolution and depletion rates of triangles and 2-simplices is the highest due their destruction by rewirings and their enhanced conversion rate via the peer pressure. They evolve faster than the edges, i.e. 1-simplices, also because any event on the edge has an effect on all the simplices that are attached to it. The node states evolve slowly to one of the single-opinion states at a quasi-stationary rate. The fast dynamics of triangles in the system is particularly important. In some parameter regimes rewiring events destroy triangles at a higher rate than it produces them. It then happens that all triangles deplete and a rewiring event can start to destroy simplices without converting triangles into simplices for simplex-conservation. We are interested in the depletion time of triangles τ. In Figure 1b) we show the average inverse depletion time $\langle 1/\tau \rangle$. It can be seen rather unsurprisingly that triangles don't deplete in the absense of rewirings. Further it can be seen that depletion rates increase as the rewiring probability increases, but less so for higher peer pressures. One may also see that τ diverges as the fragmentation transition is approached. We can explain these curves heuristically: The stronger the fragmentation, i.e. the lower the density of active links ρ, the larger become the mean degrees $\tilde{\mu}$ of the respective communities

$$\tilde{\mu} \approx \mu(1-\rho) \, .$$

Thus, rewired active links are more likely to create new triangles in any of the communities and more unlikely to destroy them due to the few inhomogeneous triangles.

Summary. We have shown, how to naturally (from the viewpoint of applications) and minimally (from the mathematical perspective) extend the co-evolving voter model to a model on simplicial complexes. It seems now plausible as further steps to also extend other adaptive contact processes to simplicial complexes, e.g., epidemic spreading models. We demonstrated that the model still yields a fragmentation transition upon varying the re-wiring rate. Yet, the quantitative properties are changed and we observe faster transitions to a single-opinion absorbing state or towards a fragmented two-opinion state. Furthermore, we found that the simplicial adaptive voter model often displays multiple time scales.

References

1. D. Stauffer. A biased review of sociophysics. *J. Stat. Phys.* 151(1-2):9–20,2013.
2. A.R. Benson, R. Abebe, M.T. Schaub, A. Jadbabaie, and J. Kleinberg. Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. USA*, 115(48):E11221–E11230, 2018.
3. F. Vazquez, V.M. Eguíluz and M. San Miguel. Generic absorbing transition in coevolution dynamics. *Phys. Rev. Lett.*, 100:108702, 2008.
4. L. Horstmeyer and C. Kuehn. An adaptive voter model on simplicial complexes. arxiv:1909.05812v1, 2019

An ego-centric view of ego-network evolution

Kleio Antoniou[1] and Demetris Antoniades[2]

[1] Open University Cyprus, Nicosia, Cyprus
kleio.antoniou@st.ouc.ac.cy,
[2] Research Center on Interactive Media, Smart Systems and Emerging Technologies (RISE), †
Nicosia, Cyprus
d.antoniades@rise.org.cy

1 Introduction

Social micro-blogging networks, like Twitter, are designed to allow their users to disseminate information and opinions with their digital peers. The information disseminated over the network characterizes the initiator but also influences it's peers perspective over them, resulting in changes in the ego network of both the initiator and the receiver. In this study, we explore how the user's activity, occupation and interests influence the evolution of her ego network. We continuously monitor individual Twitter users for a period of one month and observe how their online activity and their general characteristics, affect their ego-network in a day-to-day basis.

Over the years, significant research has shown that the total interactions between individuals in society lead to the development of complex community structures in a social network [3, 5, 7, 17, 18], composed of well-connected circles of friends, families or professional cliques [11, 13, 15] Additionally, because of the frequent changes in the patterns of activity and communication of individuals, the relevant social and communication networks are constantly under development [4, 6, 8, 16]

In recent decades, interdisciplinary network research has explored the structural and evolutionary qualities of online social graphs and the communities they include, revealing universal patterns of their dynamics. [2, 9, 12, 14, 19]

Research into the development of the ego network proves a person's connectivity, and activity is widely distributed [12], The number of edges in a social network grows as the number of nodes increases, and the average path length is shrunk by the addition of new nodes [10] after an initial extension phase [1].

In this study we examine how the characteristics and activity of the ego affect her ego network evolution. To investigate the ego network evolution in social networks, we followed 1,000 Twitter users for a period of 30 days, collecting a snapshot of their ego network every day. We categorized the users in nine professional classes [3], according to the users stated profession, as well as a random sample class. For each class we selected

†This project has received funding for the European Unions Horizon 2020 research and innovation programme under grant agreement No 739578 and the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

[3]The classes we studied were: Athletes, Politicians, Doctors, Journalists, Lawyers, Business Owners, Actors, Models and Singers

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

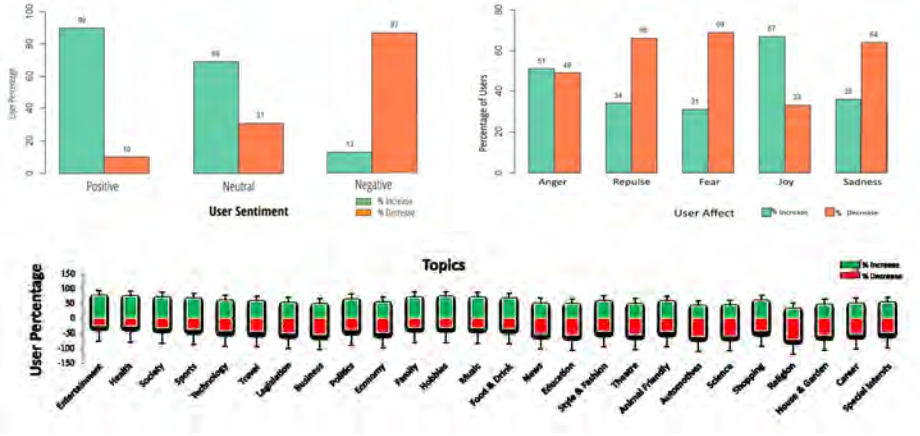


Fig. 1. Ego exhibited Sentiment, Affect and Discussion Topic effect on ego-network evolution

10 users with pre-specified initial network sizes (from 100 to 10,000 followers/friends each class). For the whole dataset, each professional class but also each initial network size we then examine the critical factors that affect the user's ego network evolution.

Our research examines how the user's ego-network changes over time but also how the characteristics of that ego-network (i.e. clustering coefficient and number of communities) evolve over the observation time period. Additionally, for each of these characteristic we examine how the activity of the user (i.e. the sentiment, opinion topics and affective tone observed in her tweets) but also the profession and initial network size affect the evolution of the ego network.

2 Results

Our initial results support our hypothesis for the role of ego characteristics, online behavior and interests in the evolution of the ego-network. Our temporal study shows daily fluctuations in the users ego-network in the range of $\pm 2\%$, equally split in increases and decreases. More than 60% of the users experience an increase in their ego-network over the period of one month, with increase rates going up to 4%. Doctors exhibit the most growth, with 90% of the category members to show increase in their network during the observation period. Additionally, we also observe a rich-get-richer phenomenon, where users with the biggest initial network (i.e. 9000-10000 followers) are the ones that observe the highest and more constant increase in their ego-network over time.

Furthermore, as depicted in 1 Twitter user's tone and subject of information disseminated plays an important role in her ego network evolution. In the upper left figure we can observe that positive sentiment (averaged over all tweets of the user in the observation period) results in an increase of the ego-network in 90% of the times, while negative sentiment results in the network decreasing 87% of the time. The upper right figure similarly shows that positive affect (i.e. Joy) can result in a ego network increase in 67% of the time. Negative affects, such as disgust, fear and sadness result in a decrease of the ego-network. Finally, the low figure shows how different topics of discussion affect the

evolution of the social network. It shows that Twitter users that mostly discuss family, hobbies and entertainment issues experience the most increase in their ego network.

Summary. Our initial analysis depicts the degree in which ego characteristics, such as sentiment, affective tone, profession, as well as the topic mostly exhibited by the user during her social networking activity, affects the evolution of her ego network topology.

References

1. Ahn, Y.Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th international conference on World Wide Web. pp. 835–844. ACM (2007)
2. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 44–54. ACM (2006)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science 286(5439), 509–512 (1999)
4. Barabâsi, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical mechanics and its applications 311(3-4), 590–614 (2002)
5. Dorogovtsev, S.N., Mendes, J.F.: Evolution of networks: From biological nets to the Internet and WWW. OUP Oxford (2013)
6. Ebel, H., Davidsen, J., Bornholdt, S.: Dynamics of social networks. Complexity 8(2), 24–27 (2002)
7. Faust, K.: Using correspondence analysis for joint displays of affiliation networks. Models and methods in social network analysis 7, 117–147 (2005)
8. Holme, P., Edling, C.R., Liljeros, F.: Structure and time evolution of an internet dating community. Social Networks 26(2), 155–174 (2004)
9. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. science 311(5757), 88–90 (2006)
10. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 177–187. ACM (2005)
11. Liljeros, F., Edling, C.R., Amaral, L.A.N., Stanley, H.E., Åberg, Y.: The web of human sexual contacts. Nature 411(6840), 907 (2001)
12. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. pp. 29–42. ACM (2007)
13. Newman, M.E., Park, J.: Why social networks are different from other types of networks. Physical review E 68(3), 036122 (2003)
14. Palla, G., Barabási, A.L., Vicsek, T.: Quantifying social group evolution. Nature 446(7136), 664 (2007)
15. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the national academy of sciences 101(9), 2658–2663 (2004)
16. Wagner, C.S., Leydesdorff, L.: Network structure, self-organization, and the growth of international collaboration in science. Research policy 34(10), 1608–1618 (2005)
17. Watts, D.J., Dodds, P.S., Newman, M.E.: Identity and search in social networks. science 296(5571), 1302–1305 (2002)

18. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. nature 393(6684), 440 (1998)
19. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM (2011)

Modularity-based selection of optimal slicing in temporal network clustering

Matteo Magnani[1], Petter Holme[2], Tsuyoshi Murata[3], and Christian Rohner[1]

[1] InfoLab, Dept. of Information Technology, Uppsala University, Sweden
{matteo.magnani,christian.rohner}@it.uu.se
[2] Institute of Innovative Research, Tokyo Institute of Technology,
Nagatsuta-cho 4259, Midori-ku, Yokohama, Kanagawa, 226-8503, Japan
holme@cns.pi.titech.ac.jp
[3] Department of Computer Science, School of Computing Tokyo Institute of Technology
W8-59 2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan
murata@c.titech.ac.jp

1 Introduction

The main approaches to cluster temporal networks include two steps: they first slice the temporal network into a sequence of static networks, then apply a clustering algorithm for multi-slice networks. However, while several methods to cluster multi-slice networks exist, assuming that the number of slices leading to a good clustering is known is typically an unrealistic assumption.

In this paper we focus on one of the best-known methods to cluster multi-slice networks: generalized Louvain [3]. Being this method based on an objective function of cluster quality (modularity), to find an optimal number of slices we might be tempted to run the generalized Louvain optimization algorithm for different numbers of slices and pick the result with the highest modularity. Unfortunately, we cannot use modularity to compare the clusterings of different slicings.

Figure 1a shows the modularity of the clusterings discovered by the generalized Louvain algorithm on four real temporal networks varying the number of slices. We can see that the more slices we have, the higher the modularity we get from the algorithm. This suggests that raising values of modularity for different numbers of slices are not necessarily an indication of better clusterings, but just a by-product of the increased size of the input networks. This is confirmed by executing the method against synthetic data where the same edges[4] are replicated on all slices. Despite introducing no new information, the modularity increases because of the addition of new edges, following a pattern that can be expressed analytically as shown in Figure 1b.

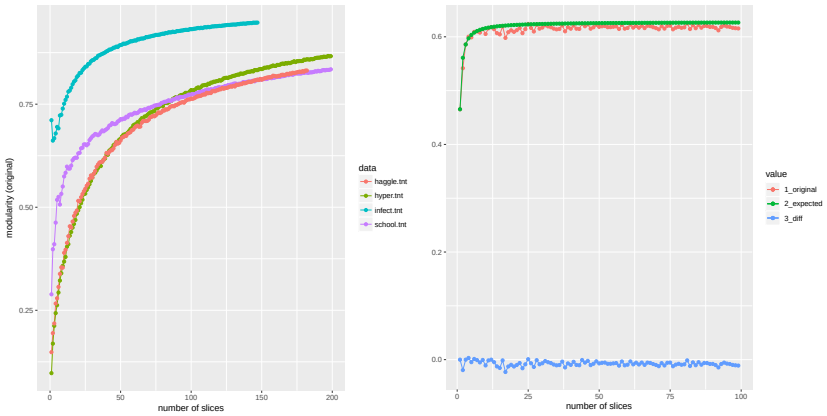
2 Method

This work is based on the assumption that multi-slice modularity has two components: one that increases with better clusterings, and one that increases just because the data

[4] Zachary's karate network

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



(a) Modularity of four real networks: empirical

(b) Modularity of synthetic network: empirical and predicted

Fig. 1: Modularity of the partitions returned by the generalized Louvain algorithm varying the number of slices for different temporal networks; the value increases with the number of slices, following a predictable pattern.

size increases, e.g., if we duplicate a slice, the same cluster extended across two slices will contain additional inter-slice edges. Therefore, to identify an optimal number of slices we can try to isolate the first component in the modularity and use it to compare clusterings computed using different numbers of slices.

To remove the effect of data size we use an edge reshuffling process that destroys the clusters in the network without affecting the degree distribution [2, 1]. For each number of slices, the Louvain algorithm is run both on the original data and on the reshuffled data where the clusters have been destroyed. The modularity on the dataset without clusters indicates the effect of the number of slices on modularity, and the difference between the two indicates the part of modularity due to the presence of clusters. We call this difference *normalized multi-slice modularity*.

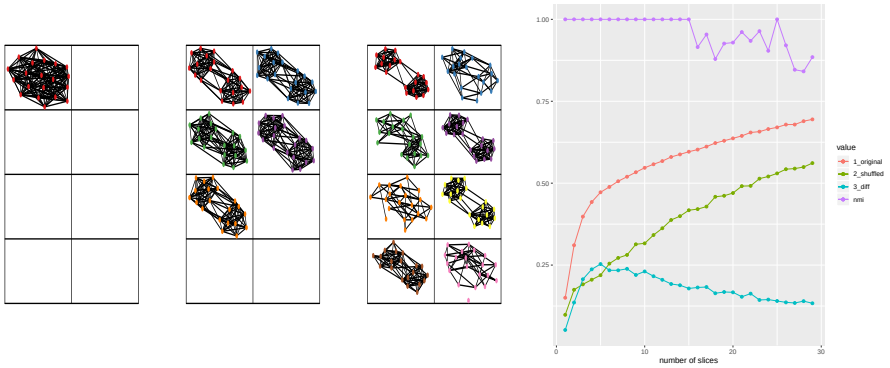
3 Results

To test our approach we built different synthetic networks where the optimal number of slices is known in advance. Here we only show one of these cases, for space reasons. This example consists of two cliques separated by random noise (20% density), with this pattern repeated five times. The network is shown in Figures 2a-Figure 2c, split into different numbers of slices. When we only have one slice, the combination of the noise present throughout the existence of the network hides the clusters. When we use five slices (Figure 2b), the cliques are easily visible in all slices. In time, the cliques disappear from some of the slices, and ultimately from all of them, because their edges get spread across several sparser and sparser slices.

With this dataset, we know that the clusters are the most visible when we have five slices. Figure 2d shows the original modularity, the randomized modularity and our

normalized multi-slice modularity. While the first two increase when the number of slices increases, the normalized multi-slice modularity has a peak at five slices.

Figure 2d also shows the normalized mutual information (NMI) between the ground truth clusters and the clusters found by the algorithm, for different numbers of slices. A higher value of NMI corresponds to more similar clusterings. We notice how the number of slices identified by our approach corresponds to the highest NMI, but the generalized Louvain algorithm would still be able to reach the same NMI with other numbers of slices (up to fifteen with this data).



(a) one slice (b) five slices (c) eight slices (d) modularity

Fig. 2: Best number of slices: controlled experiment

Summary. We propose and evaluate a method to identify an optimal number of slices based on modularity. Our work includes additional results and discussions not presented here for space reasons: Practical details on how to correctly perform edge shuffling. The application of the method to several real datasets for which no ground truth is available. A critical analysis of the method, identifying aspects that require further validation. A critical analysis of modularity-based approaches in the context of temporal network clustering, identifying scenarios such as recurrent temporal clusters that are not captured by this objective function.

Acknowledgments This work was partly funded by STINT project IB2017-6990.

References

1. Gauvin, L., Génois, M., Karsai, M., Kivelä, M., Takaguchi, T., Valdano, E., Vestergaard, C.L.: Randomized reference models for temporal network. arXiv:1806.04032v1 (2018)
2. Karsai, M., Kivelä, M., Pan, R., Kaski, K., Kerté, J., Barabási, A.L., Saramäki, J.: Small but slow world: How network topology and burstiness slow down spreading. Physical Review 83 (2011)
3. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. Science (New York, N.Y.) 328(5980), 876–8 (may 2010), http://www.sciencemag.org/content/328/5980/876.abstract

Indetermination of networks structure from the dynamics perspective

Malbor Asllani[1], Bruno Requião da Cunha[1,2] Ernesto Estrada[3,4], and James P. Gleeson[1]

[1] MACSI, Department of Mathematics & Statistics, University of Limerick, V94 T9PX Limerick, Ireland,
malbor.asllani@ul.ie,
[2] Rio Grande do Sul Superintendency, Brazilian Federal Police, Av. Ipiranga 1365, 90160-093 Porto Alegre, RS, Brazil
[3] Institute of Mathematics and Applications (IUMA), Universidad de Zaragoza, Pedro Cerbuna 12, E-50009 Zaragoza, Spain
[4] ARAID Foundation, Government of Aragón, Zaragoza 50018, Spain

1 Introduction

Networks constitute a paradigm of complexity in real life systems by assembling the structure of the interactions of their elementary constituents [1, 2]. They are found at every level of biological organisation, from genes inside the cells [3] to the trophic relations between species in large ecosystems [4]. Nowadays, with the enormous development of data science, there is a huge interest related to the network inference, namely detecting the interacting structure from external measurements or observations. For example, reconstructing the structure of brain networks from the activity of neuronal patches has been a major goal in computational neuroscience [5]. The dynamics that takes place on networked systems can, in some cases, strongly influence the perception that we have regarding local topological features such as the degree [6] or global ones such as network non-normality [7].

In this work, we focus specifically on the problem of measuring network centralities from the dynamical point of view. We show that the inference of networks' structural properties depends heavily on the competition between the node-based dynamics on one hand and the interactions between the nodes on the other. In particular, we illustrate such a phenomenon based on the communicability centrality [8], considered as a reliable measure for dynamical inference [9]. We show that when the local intra-nodes dynamics is slower than the inter-nodes one then the ranking of the nodes according to the standard definition of the communicability, becomes inadequate. Such ranking can be enhanced if further information regarding the nature of the dynamics occurring on the network is available. As an example, we show that for networks with different time-scale structures such as strong modularity, the existence of fast global dynamics can imply that precise inference of the community structure is impossible.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

To illustrate our analysis we will consider the *SI* model for epidemic spreading in a metapopulation network [10, 11]. Such a formulation of the spreading processes has been employed to model, for example, the propagation of misfolded proteins in neurodegenerative diseases [12]. The mean-field dynamics reads:

$$\begin{aligned} \dot{S}_i &= -\alpha S_i I_i + (1-\alpha)\sum_j \mathscr{L}_{ij} S_j \\ \dot{I}_i &= \alpha S_i I_i + (1-\alpha)\sum_j \mathscr{L}_{ij} I_j \,, \end{aligned} \tag{1}$$

where S, I are the concentrations, respectively, of the susceptible and the infected individuals, α is the infection rate, $1-\alpha$ the diffusion constant and $\mathscr{L}$ is the Laplacian matrix defined as $\mathscr{L}_{ij} = \mathscr{A}_{ij} - k_i$ where k_i is the degree of node i [2]. Starting from this model, we will compare the effectiveness of measuring the nodes' centrality from the dynamical observables and compare it to different structural definitions (communicability, modularity etc). To do so we first select the most central node of the graph (e.g., the one with the highest betweenness) as the observation node and then take the time needed for the infection to reach such node as the dynamical observable. We will indicate the observable as RT_i and will refer to it as the corresponding *reaching time* for the starting node i. We prove that if the dynamics of the network outcompetes that of the nodes, $\alpha < 1/2$ then the range of values taken by the reaching time RT_i over all nodes i is small. This means in the presence of noise in the experimental data (due to the stochastic nature of the process and measurements) it is not possible to distinguish the nodes anymore. To emphasize this point, we consider a strongly modular topology [2], a feature of crucial importance in modern computational neuroscience [5]. In Fig. 1 we show that in a general system which dynamics depends on both the network connections and the node dynamics as in eqs. (1) is not possible to infer the structural properties as e.g. the modularity in a correct way. Moreover, the accuracy of the resolution depends on the competition between these two dynamical components of the system.

Summary: In order to determine the role that each node has inside a complex network, several centrality measures have been developed so far in the literature. In this paper, we show that when the dynamics taking place at the local level of the node is slower than the global one between the nodes, then the system may lose track of the structural features. On the contrary, when that ratio is reversed only global properties such as the shortest distances can be recovered. In this sense, our results constitute an uncertainty principle where inferring the structural properties of a network at a global level (e.g. modularity) means sacrificing resolution of the local dynamics of the nodes, and vice-versa. For illustration purposes, we show that for strong modular networks, the existence of fast global dynamics can imply that precise inference of the community structure is impossible, particularly in the presence of noise.

References

1. E. Estrada, *The structure of complex networks: theory and applications* , Oxford University Press (2012).

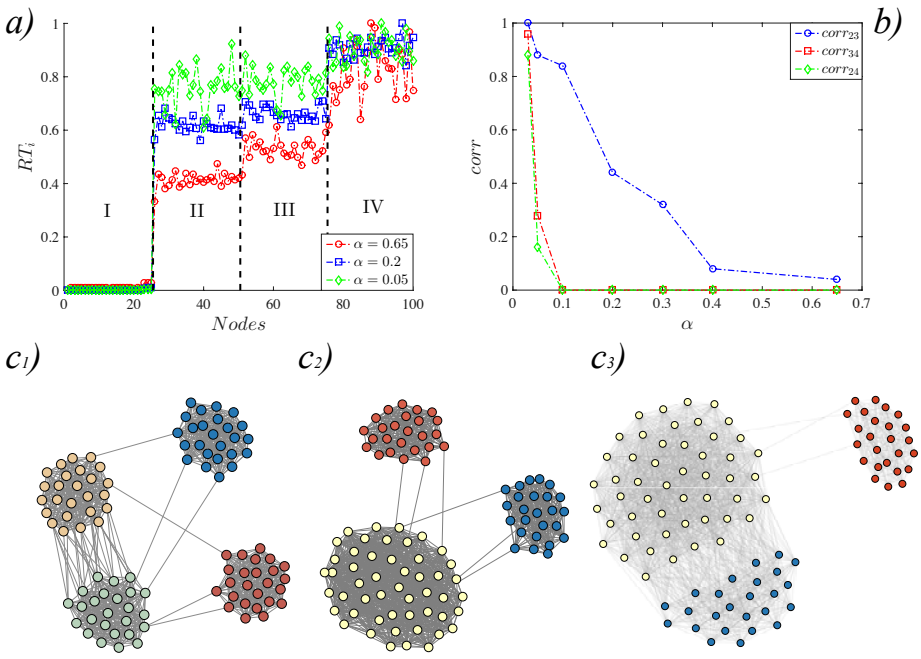


Fig. 1. *a*) We plot the normalised reaching time RT_i variable of the four modules (indicated by roman numbers) showing that for decreasing values of the α parameter (as in the legend) the ranges of the dynamical variables for different modules overlap. *b*) The correlation variable for each couple of modules (with the exception of the first) as a function of α. *c*) We show how the resolution of a given reconstruction method can be affected by different choices of the tuning parameter (for the same values as in panel *a*)). *c*) A representative visualisation of the networks reconstruction where it is shown the gradual deformation perceived in the network modularity from: c_1) ($\alpha = 0.65$) the original 4 modules topology, c_2 ($\alpha = 0.2$) modules *II* and *III* have merged and c_3 ($\alpha = 0.05$) where module *IV* is now merging with the union of the modules $II - III$. The modular network has 100 nodes and has been generated through a Stochastic Block Model with total link density $p = 0.2$ and probability 0.01 for an inter-module link.

2. M. E. J. Newman, *Networks: An introduction* 2ed, Oxford University Press (2018).
3. S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Nat. Gen. **31**(1), 64, (2002).
4. D. Garlaschelli, G. Caldarelli and L. Pietronero, Nature **423**(6936), 165, (2003).
5. O. Sporns, *Networks of the Brain*, MIT Press (2010).
6. M. Asllani, T. Carletti, F. Di Patti, D. Fanelli, F. Piazza, Phys. Rev. Lett. **120**(15) 158301 (2018).
7. M. Asllani, R. Lambiotte and T. Carletti, Sci. Adv. **4** eaau9403 (2018).
8. E. Estrada and N. Hatano, Phys. Rev. E **77**(3), 036111 (2008).
9. M. Gilson, N. E. Kouvaris, G. Deco, and G. Zamora-López, Phys. Rev. E **97**, 052301 (2018).
10. J. Murray, *Mathematical Biology: I. An introduction* 3ed, Springer (2002).
11. D. Brockmann and D. Heilbing, Science **342**(6164), 1337–1342 (2013).
12. Y. Iturria-Medina, R. C. Sotero, P. J. Toussaint, A. C. Evans and the Alzheimers Disease Neuroimaging Initiative, PLOS Comput. Biol. **10**(11) e1003956 (2014).

Part V

Human Behaviour

Co-evolutionary Opinion Dynamics on Adaptive Social Networks: the Role of Social Balance

Tuan Pham[1,2], Rudolf Hanel[1,2], and Stefan Thurner[1,2,3,4]
tuan.pham@meduniwien.ac.at

[1] Section for the Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Vienna, Austria,
[2] Complexity Science Hub Vienna, Josefstadterstrasse 39, A-1090 Vienna, Austria
[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[4] IIASA, Schlossplatz 1, 2361 Laxenburg, Austria

1 Introduction

The understanding of the variability and susceptibility of individuals' attitudes and opinions, when exposed to random influences from their milieu, is a central question in the social sciences. We here address this question by a spin-model of agent opinions that are coevolving with their social network. We show how groups of agents with opposing opinions form in the low exposure regime, while groups disappear above a critical level of exposure. Within the presented approach, the effect of "social balance" – a concept first introduced by Heider [1], is explicitly taken into account with a new term in the model Hamiltonian. The dynamics can be understood by the phase diagrams of the model.

2 Results

In Heider's social balance theory, a group of three individuals forms a balanced triangle, if either all the three are mutual friends or two of them are friends who both have the same enemy as the third. They form an unbalanced triad, if either all three are mutually hostile, or one of them has two friends who detest each other (see Figure. 1). If such a situation occurs, agents strive to reduce the tension in their unbalanced triangles by flipping one of the three links, so that balanced triangles tend to remain in the network. Assuming that this fact, as well as the tendency of agents to avoid contention with their neighbours, are key driving forces in social dynamics, we arrive at the Hamiltonian:

$$H = -\sum_{(i,j)} J_{ij} s_i s_j - g \sum_{(i,j,k)} J_{ij} J_{jk} J_{ki} \,, \tag{1}$$

where both the opinion s_i of individual i and the links J_{ij} between individuals i and j can take values $\{-1,1\}$. That is, opinions s_i can be yes/no answers to a political question, while links J_{ij} represent friendship and enmity relationships, respectively. In Eq. (1), the first term biases friends to be of the same opinion and enemies to be of opposing opinions, while the second term, which takes into account the effect of triangles on the system dynamics, biases triads towards "social balance". The parameter $g \in (0,1)$,

allows us to continuously control the relative weight of the topological effect. Based on this Hamiltonian, the coevolution of opinions and network links is implemented by using the Metropolis algorithm [2]. Here, at every time step, both an opinion and a tie, are chosen at random to be independently and subsequently flipped if this decreases the Hamiltonian energy H or with a probability $e^{-H/T}$ if this is not the case, where T is the social temperature which represents random influences from the individual milieu. For simplicity, we consider only fully-connected undirected networks, where every one knows everyone else.

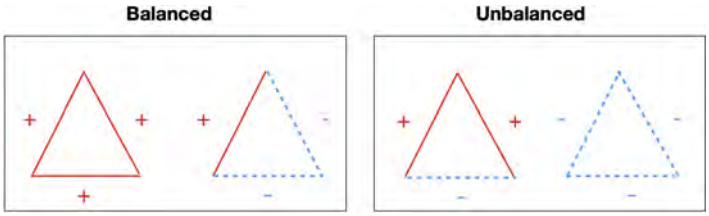


Fig. 1. Balanced and unbalanced triangles.

The network structure that is relevant to our purpose can be characterised by a topological variable f, which measures the difference of the fractions of balanced and unbalance triangles in the network:

$$f = \frac{n_{\Delta_+} - n_{\Delta_-}}{M} , \tag{2}$$

where $M = n_{\Delta_+} + n_{\Delta_-}$ and n_{Δ_+} (n_{Δ_-}) are the total number of triangles and the number of balanced (unbalanced) triangles in the network of social ties. For fully-connected networks of size N, $M = \binom{N}{3}$. Thus, $f = 1$, if all the triangles are balanced and $f < 1$ if on or more unbalanced triangle are present. We call f the "net balance".

When there is no unbalanced triangle in the network, it has a special structural property. According to Harary's theorem [3], the set of nodes is partitioned into two disjoint subsets $\mathscr{B}_1$ and $\mathscr{B}_2$, one of which may be empty, such that all links between nodes of the same subset are positive and all links between nodes of the two different subsets are negative. The existence of these two clusters suggests the definition of another measure, that we call the group difference, which characterizes the final distribution of agent opinions between them,

$$m_g = \frac{1}{N} \left\langle \left| \sum_{i \in \mathscr{B}_1} s_i - \sum_{i \in \mathscr{B}_2} s_i \right| \right\rangle . \tag{3}$$

By definition, $m_g \in [0,1]$. m_g takes its maximum value 1 if and only if each of the two clusters $\mathscr{B}_1$ and $\mathscr{B}_2$ consist of like-minded agents but the opinions are contradictory between agents belonging to different groups. This picture is analogous to what happens if two clusters of classical Ising-spins are coupled to each other by anti-ferromagnetic interactions. At low temperature, due to the ferromagnetic interactions between spins

inside a cluster, they are aligned in the same direction, but spins in different clusters must have opposite directions as their interactions are anti-ferromagnetic.

In figure 2, we show that by Heider's structural balance, the society can eventually reach a balanced state in which opinions are split into two disjoint groups respecting this principle (yellow region in the figure). The stronger the effect of triangles is (i.e., the larger g), the more stable this bi-partition is against the destructive effect of the social temperature T. However, for fixed g, as long as the temperature increases, these clusters disappear and opinions become randomly distributed amongst agents (the dark blue region in the figure), marking a continuous phase transition in both the net balance, f, and the group difference, m_g.

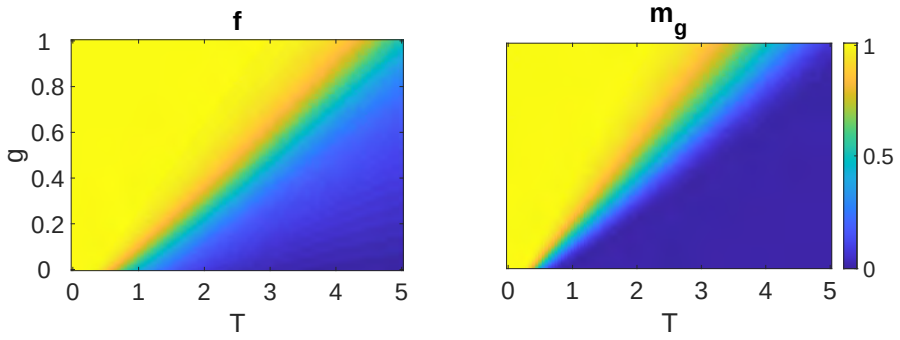


Fig. 2. The net balance f (left) and the group difference m_g (right), as a function of the social temperature T and the relative strength of the triangle effect compared to the agent pair-wise interaction g. Results averaged over 10^3 realizations of the model (1) by the Metropolis algorithm with 10^4 time steps for fully-connected networks with $N = 10$ nodes.

3 Summary.

We investigated the role of social balance in the coevolution of individual opinions and their social network. In particular, we have shown how this effect can lead to a simple understanding of the polarization of society that is observed today. Within the new framework, the question about the stability of this polarization under social perturbations can also be fully addressed.

References

1. Heider, F.: Attitudes and Cognitive Organization. Journal of Psychology 21, 107 (1946)
2. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E.: Equation of state calculations by fast computing machines. The Journal of Chemical Physics, vol. 21, no. 6, 1087–1092 (1953)
3. Harary, F. : On the notion of balance of a signed graph. Michigan Math. J., vol. 2, no. 2, 143 –146 (1953)

Similarity forces and recurrent components in face-to-face interaction networks

Marco Antonio Rodríguez Flores[1] and Fragkiskos Papadopoulos[1]

Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, 33 Saripolou Street 3036 Limassol, Cyprus
mj.rodriguezflores@edu.cut.ac.cy

1 Introduction

Understanding the dynamics of face-to-face interaction networks is essential for expanding our knowledge of how diseases spread, how information is exchanged or how communities form and evolve [1]. However, it has been difficult to find simple processes that reproduce the structural and dynamical properties of these networks including the recurrent formation of groups of the same people, which originate from human motion patterns that are far from random [2]. For this reason, models like the attractiveness model [3] that are based on mobile interacting agents that perform random walks are unable to reproduce the abundance of recurrent components found in real systems, even thought they can reproduce a variety of other important properties.

In this extended abstract we present the *Force-directed Motion* model (FDM), which has been recently published in PRL [4]. The model suggests that hidden similarity distances between the agents act as forces that direct their motion towards each other in the physical space where they move, and determine the duration of their interactions. The FDM reproduces a wide range of properties of real systems, including the formation of recurrent components.

2 Model description

The FDM assumes that the agents move and interact in a closed two-dimensional Euclidean space (an $L \times L$ square), and that they also reside in a hidden similarity space. Our choice for the similarity space is the simplest metric space, a circle of radius $R = N/2\pi$ where each agent $i = 1,2,...,N$ is assigned a random angular coordinate $\theta_i \in [0,2\pi]$. Thus, the similarity distance between two agents i, j is $s_{ij} = R\Delta\theta_{ij}$, where $\Delta\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$ is the angular distance between them.

Time in the model is slotted and at the beginning of each slot $t = 1,2,\ldots,T$ the agents can be either *inactive* or *interacting*. Then:

1. Each inactive agent i is activated with a preassigned probability r_i.
2. Each interacting agent i escapes (i.e., quits) its interactions with probability:

$$P_i^e(t) = 1 - \frac{1}{|\mathcal{N}_i(t)|} \sum_{j \in \mathcal{N}_i(t)} e^{-s_{ij}/\mu_1}, \tag{1}$$

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where $\mathcal{N}_i(t)$ is the set of agents that are interacting with i in slot t, while parameter μ_1 is the decay constant allowing us to control the average contact duration.

3. Each agent i that becomes active or escapes its interactions updates its position $\mathbf{q}_i^t = (x_i^t, y_i^t)$ according to the following motion equation:

$$\mathbf{q}_i^{t+1} = \mathbf{q}_i^t + \sum_{j \in \mathcal{S}(t)} F_{ij} \frac{(\mathbf{q}_j^t - \mathbf{q}_i^t)}{||\mathbf{q}_j^t - \mathbf{q}_i^t||} + \mathbf{v}_i, \tag{2}$$

where $\mathcal{S}(t)$ is the set of all moving and interacting agents in the slot, $\mathbf{v}_i = (v\cos\phi_i, v\sin\phi_i)$ is the random motion component, where ϕ_i is sampled uniformly at random from $[0, 2\pi]$ and $v \geq 0$ is the random displacement magnitude. $F_{ij} = F_0 e^{-s_{ij}/\mu_2}$ is the magnitude of the *attractive force* between agents i and j. Parameters F_0 and μ_2 control the rate at which recurrent components form as well as the size of the largest component.

4. All agents that updated their positions transition to the interacting state if they are within interaction range d from other non-inactive agents. Otherwise, they transition to the inactive state.

3 Results and discussion

As an illustrative example here, we use the FDM to model the face-to-face interaction network of a Primary School in Lyon, France [5]. This temporal network consists of the interactions between 242 individuals over 2 days for approximately 8.5 hours in each day. Interactions were registered every 20 seconds if the individuals were facing each other within a range of 1-1.5 meters. The total number of non-empty snapshots of 20 seconds in the data is 3100. However, we remove the snapshots corresponding to the lunch break period in each day when some students go home to eat and the others interact in the common grounds of the school. This leaves us with 2378 snapshots.

We generate an FDM temporal network with parameters: $N = 242$, $T = 2378$, $L = 98$, $\mu_1 = 0.35$, $F_0 = 0.2$, $\mu_2 = 0.78$, $v = d = 1$ and $r_i = 0.5$ for each agent i (details of how to tune the model parameters in the modeled counterparts of real systems can be found in [4]). For comparison we also generate a temporal network with the attractiveness model [3], with parameters $N = 242$, $T = 2378$, $L = 50$, $v = d = 1$ and $r_i = 0.5$ for each agent i.

Fig. 1a shows the recurrent components observed during the first day in the Primary School with the observation period (x-axis) binned into intervals of 30 minutes. Figs. 1b,c correspond to the recurrent components observed in simulated networks with the attractiveness model and the FDM. To generate these plots, we have extracted the unique components found in the respective network and assigned them IDs in order of appearance. In the plots, the recurrent components, i.e., the components that appeared at least once in a previous time interval are marked with blue lines. The recurrent components in the FDM are as abundant as in the real network. In stark contrast, they are scarce in the attractiveness model. In Figs. 1d,e we also see that the FDM reproduces other important properties of the real network, like the distributions of the contact and intercontact durations. Finally, Fig. 1f shows the probability that two agents are connected in a slot as a function of their similarity distance s_{ij} in the FDM. Remarkably,

without enforcing it into the model, this probability resembles the Fermi-Dirac connection probability in the $\mathbb{S}^1$ model of non-mobile complex networks [6]. We explore this connection in [7].

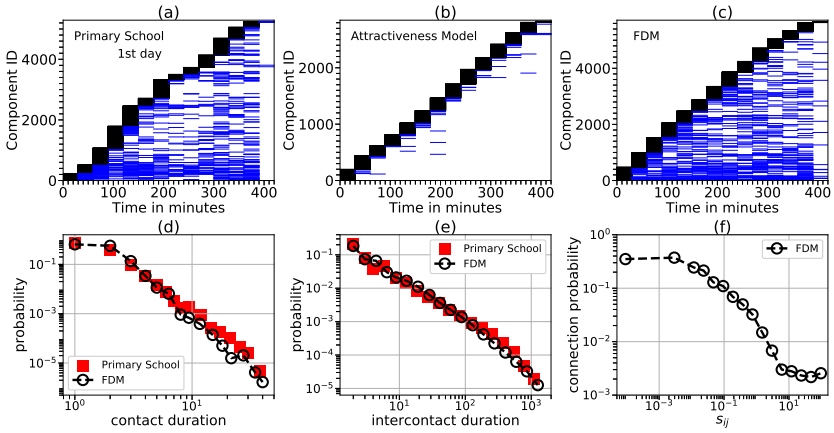


Fig. 1. Top row: Unique and recurrent components in the Primary School (a), a simulated network with the attractiveness model (b), and a simulated network with the FDM (c). Bottom row: distributions of contact (d) and intercontact (e) durations in the Primary School and the FDM. (f) Probability that two agents are connected in a slot as a function of their similarity distance s_{ij} in the FDM.

We report similar results for other face-to-face interaction networks and illustrate a similar behavior of spreading processes in real and FDM-simulated networks in [4]. Our results pave the way towards simple yet realistic models of face-to-face interaction networks.

References

1. Barrat, A., Cattuto, C.: Face-to-face interactions. Social Phenomena: From Data Analysis to Models (2015)
2. Sekara, V., Stopczynski, A., Lehmann, S.: Fundamental structures of dynamic social networks. Proc. Natl. Acad. Sci. 113(36), 9977 – 9982 (2016)
3. Starnini, M., Baronchelli, A., Pastor-Satorras, R.: Model reproduces individual, group and collective dynamics of human contact networks. Social Networks. 47, 130 – 137 (2016)
4. Rodríguez-Flores, M.A., Papadopoulos, F.: Similarity forces and recurrent components in human face-to-face interaction networks. Phys. Rev. Lett. 121, 258301 (2018)
5. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.F, Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., Vanhems, P.: High-resolution measurements of face-to-face contact patterns in a primary school. PLoS ONE. 6(8), e23176 (2011)
6. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., Boguñá, M.: Hyperbolic geometry of complex networks. Phys. Rev. E. 82(3), 036106 (2010)
7. Papadopoulos, F., Rodríguez-Flores, M.A.: Latent geometry and dynamics of proximity networks. arXiv:1907.00073 (2019)

Detecting eigenmoods in individual human emotions.

Marijn ten Thij[1], Johan Bollen[1,2], and Luis M. Rocha[1,3]

[1] Center for Social and Biomedical Complexity, Indiana University, Bloomington, IN, USA
[2] Wageningen University, Wageningen, the Netherlands
[3] Instituto Gulbenkian de Ciência, Oeiras, Portugal

1 Introduction

Social media platforms record a multitude of information pertaining to the behavior and language of billions of individuals. Emotions play a crucial role in these phenomena but are rarely explicitly expressed [2]. They must therefore be assessed from text content by sentiment analysis algorithms. However, the high frequencies of common terms in a language can obscure actual expressions of sentiment. For example, the positive sentiment values of holiday greetings (e.g. "happy holidays") will bias many sentiment analysis tools towards positive assessment regardless of actual sentiment fluctuations. This same effect may obscure the diverging emotional responses of sub-populations, e.g. in the case of significant sports events or elections (e.g. "win" vs "lose"). A similar issue may occur in the case where individual sentiment fluctuates simultaneously along different dimensions or instances of mood, such as Valence and Arousal, or Activation [3, 6, 1].

2 Results

Following [8], we leverage the Singular Value Decomposition (SVD) [4] of a sentiment-time matrix to separate actual changes in user sentiment from sentiment observations resulting from default term frequencies in a language. In effect, we show that the SVD reveals "eigenmood" from sentiment analysis data by their decomposition into singular value approximations.

We demonstrate this approach using a sample of 3,624 *Twitter* users that mentioned a mental health issue such as depression in at least 1 tweet. We obtained their individual timelines, i.e. a longitudinal record of their most recent 3,200 messages, from the Twitter API. We estimate a tweet's Valence, Arousal, and Dominance sentiment from the average CRR ANEW lexicon [5] ratings of its terms. From these scores, we create a time-series of weekly averaged sentiment scores for each individual user.

Aggregating these time-series for all users we obtain a probability distribution of mean sentiment values for each week in our data. This results in a matrix of weekly sentiment distributions which we use as the basis of our analysis. For all users we consider sentiment values for a time span of 80 weeks, i.e. January 2nd 2017 through July 15th 2018. The resulting matrices are visualized in Fig. 1 A and E as heat maps in which the color intensity of each cell indicates the number of tweets whose sentiment value falls in a given sentiment bin.

COMPLEX NETWORKS 2019

The 8th International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The SVD factorizes a matrix M in three matrices $U \cdot \Sigma \cdot V$ where the matrix Σ contains the singular values of the matrix M. Our approach isolates distinct eigenmoods from these singular values, the distribution of which is shown in Fig. 1 D. The largest singular value has a disproportionate magnitude earlier shown to correspond to the base sentiment distribution of the English language [8, 7].

We can construct different approximations of M or remove noise by retaining singular vectors of interest. For instance, if we only retain the first singular value in the top-left spot of a matrix $\tilde{\Sigma}$ (by setting every other entry in the diagonal matrix to 0) and compute $U \cdot \tilde{\Sigma} \cdot V$, we obtain an approximation $\tilde{M}^1$ of M shown in Fig. 1 B and F. These reconstructed matrices capture the expected stable sentiment distribution of the English language. In contrast, if we remove the first singular vector, by calculating $M - \tilde{M}^1$, we obtain the matrices shown in Fig. 1 C and G. In Fig. 1 we observe a bi-modal sentiment distribution in our sample group (two yellow bands in Fig. 1 C), ending approximately at week 50, which was previously hidden in the overall sentiment distribution captured by $\tilde{M}^1$. We obtain similar but visually less pronounced effects when applying this technique to the longitudinal sentiment of single individuals (an example shown in Fig. 1 E, F and G).

The detection of eigenmoods in aggregate or individual social media sentiment may enable the characterization of change points by projecting the sentiment distribution of individual weeks along different singular vectors of our decomposition as previously demonstrated by [8]. This approach may have applications to the detection of changes in individual sentiment related to the dynamics of mood disorders.

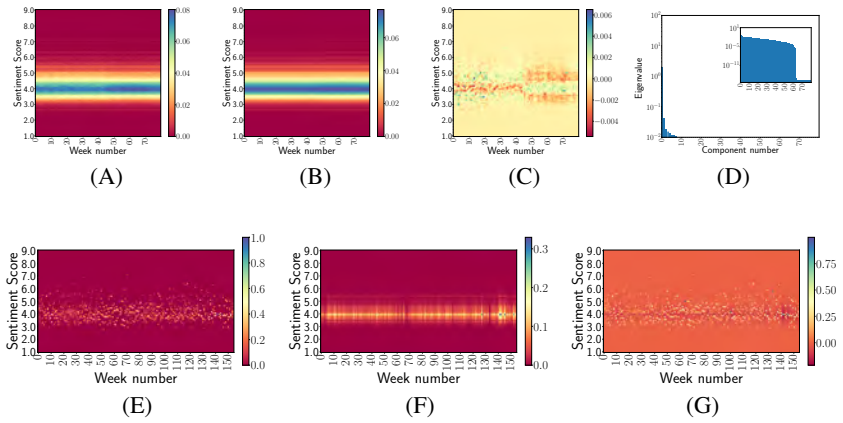


Fig. 1: Eigenmood analysis of Twitter sentiment distributions. **A** and **E**: mood matrix (M) for a group of users and a randomly chosen individual respectively. **B** and **F**: first singular value approximation ($\tilde{M}^1$). **C** and **G**: remaining sentiment signal after removal of first singular value approximation from original $M - \tilde{M}^1$). **D**: spectrum of singular values for group sentiment-time matrix.

References

1. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences (2017), https://www.pnas.org/content/early/2017/08/30/1702247114
2. Fan, R., Varol, O., Varamesh, A., Barron, A., van de Leemput, I.A., Scheffer, M., Bollen, J.: The minute-scale dynamics of online emotions reveal the effects of affect labeling. Nature Human Behaviour 3(1), 92–100 (2019), https://doi.org/10.1038/s41562-018-0490-5
3. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dvelopmental Psychopathology 17, 715–734 (2005)
4. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: A practical approach to microarray data analysis, pp. 91–109. Springer (2003)
5. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior Research Methods 45(4), 1191–1207 (Dec 2013), https://doi.org/10.3758/s13428-012-0314-x
6. Watson, D., Tellegen, A.: Toward a consensual structure of mood. Psychological Bulletin 98, 219–235 (1985)
7. Wood, I.B., Gonalves-S, J., Bollen, J., Rocha, L.M.: Measuring collective mood variation. (In Preparation)
8. Wood, I.B., Varela, P.L., Bollen, J., Rocha, L.M., Gonçalves-Sá, J.: Human Sexual Cycles are Driven by Culture and Match Collective Moods. Scientific Reports 7(1), 17973 (2017), https://doi.org/10.1038/s41598-017-18262-5

The Language of Peace is Complex

Luca Maria Aiello

Nokia Bell Labs, Cambridge, UK
lajello@gmail.com,
WWW home page: http://www.lajello.com

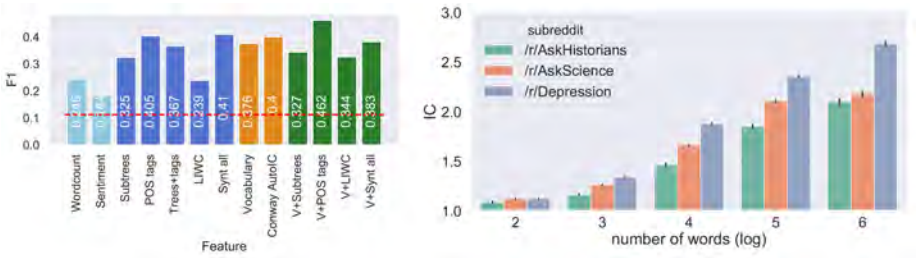


Fig. 1. Left: Classifier performance with different feature sets: word count, sentiment, syntactic features (in blue), vocabulary-based features (in orange), and combinations of them (in green). Vocabulary plus POS Tags is the best performing approach. The red line shows the random baseline. **Right**: Average IC of texts in the three subreddits considered, binned by text length (log of the number of words). Depression-related posts and comments have higher IC compared to texts of comparable length from the other two subreddits.

Social networks are heavily polarized [5,3], which calls for technological solutions that can effectively bridge conflicting communities. In the past, researchers have studied conflict on social media and its effect on the network structure as well as on the use of language [4,9]; however, it is still unclear what are the best strategies to *resolve* conflict. We propose a computational social science solution to the problem of conflict resolution by operationalizing the concept of Integrative Complexity.

Integrative Complexity (IC) is a psychometric that measures the ability of a person to recognize multiple perspectives on a particular issue and connect them, thus identifying paths for conflict resolution [13]. The lowest end of the IC spectrum is associated with inflexible, fixed perspective thinking and the highest end with integrating groups of perspectives in an elaborate, hierarchical fashion [2]. IC has been applied to a wide range of source materials, including diplomatic communications, political speeches, personal correspondence and legal judgments [13]. As a result, it has been presented as a powerful predictor for a variety of outcomes, such as international conflict [12], aggression [14] and political preferences [7]. However, scoring the IC of a text is a manual, time-consuming task to be carried out by trained experts. Previous efforts have attempted to automatizing IC scoring with simple vocabulary-based classifiers [6,1]. However, in its original definition, IC is concerned not with *what* we say, but *how* we say it. In this work [10], we show that syntactic information is crucial to generalize automated IC scoring.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

From an extensive corpus of text manually labeled with IC scores [6], we extract several families of textual features (text length, POS Tags, Dependency subtrees, LIWC, sentiment) and use them to classify the level of IC in documents. The combination between vocabulary features and syntax features (POS Tags) outperforms all previous approaches and other feature combinations (Figure 1 left).

We run for the first time a large-scale analysis of Integrative Complexity expressed in social media by applying our model to 400k+ Reddit posts, with the goal of building evidence about our method's external validity. We based our analysis on previous literature [11] that showed that the level of IC tends to increase during periods of severe personal distress (e.g., following the death of a loved one or a betrayal). We therefore compare texts from /r/depression, a forum intended for sharing negative experiences and providing social support, with other two communities, /r/AskScience and /r/AskHistorians, which are focused on knowledge exchange. In agreement with the theory, we find that posts in the /r/depression subreddit, where users write about their experience of depression often triggered by difficult personal circumstances, grief, and other traumas [8], exhibit higher IC than what is measured in the discussions about non-dysphoric experiences of the other two fora (Figure 1 right). We provide extensive quantitative and qualitative analysis of the posts to support our findings.

From the theoretical standpoint, this work reinforces the evidence that IC can be effectively operationalized and that it can be done most effectively when language syntax is brought into the equation. By opening our method to the research community, we hope to encourage its application to a wider range of domains; in particular, we believe it can enable important practical applications in social media analytics. Since previous research has shown that Integrative Complexity is a good predictor of the richness of dialogue [7], we believe that automatic measurement of IC will have an important role in tackling the resolution of conflicts in an increasingly polarized social media space.

References

1. A. K. Ambili and K. M. Rasheed. Automated scoring of the level of integrative complexity from text using machine learning. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 300–305. IEEE, 2014.
2. G. Baker-Brown, E. J. Ballard, S. Bluck, B. De Vries, P. Suedfeld, and P. E. Tetlock. Coding manual for conceptual/integrative complexity. *Berkeley, CA: University of British Columbia and University of California*, 1990.
3. E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
4. J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Icwsm*, pages 61–70, 2015.
5. M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.
6. L. G. Conway, K. R. Conway, L. J. Gornick, and S. C. Houck. Automated integrative complexity. *Political Psychology*, 35(5):603–624, 2014.
7. L. G. Conway, L. J. Gornick, S. C. Houck, C. Anderson, J. Stockert, D. Sessoms, and K. McCue. Are conservatives really more simple-minded than liberals? the domain specificity of complex thinking. *Political Psychology*, 37(6):777–798, 2016.

8. M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*, 2014.
9. K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):3, 2018.
10. A. Robertson, L. M. Aiello, and D. Quercia. The language of dialogue is complex. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 428–439, 2019.
11. P. Suedfeld and S. Bluck. Changes in integrative complexity accompanying significant life events: Historical evidence. *Journal of personality and Social Psychology*, 64(1):124, 1993.
12. P. Suedfeld, P. E. Tetlock, and C. Ramirez. War, peace, and integrative complexity: Un speeches on the middle east problem, 1947–1976. *Journal of Conflict Resolution*, 21(3):427–442, 1977.
13. P. Suedfeld, P. E. Tetlock, and S. Streufert. Conceptual/integrative complexity. In C. P. Smith, editor, *Motivation and personality: Handbook of thematic content analysis*, pages 393–400. Cambridge University Press, 1992.
14. D. A. Winter. Slot rattling from law enforcement to lawbreaking: A personal construct theory exploration of police stress. *International journal of personal construct psychology*, 6(3):253–267, 1993.

Comparing gender mixing preferences across networks

Leto Peel[1], Mauro Faccin[1], Fariba Karimi[2], and Matteo Cinelli[3]

[1] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université Catholique de Louvain, Louvain-la-Neuve B-1348, Belgium,
[2] Department of Computational Social Science, GESIS, Cologne, Germany,
[3] Applico Lab, CNR-ISC, 00185, Rome, Italy

Assortative mixing in networks is the tendency for nodes with the same attributes, or metadata, to link to each other. It is a property often found in social networks manifesting as a higher tendency of links occurring between people with the same age, race, or political belief. Assortativity by gender has often found to be weak or non-existent (e.g. [7]) when measuring it as a global average across a population. However, recent work has demonstrated that more specific gender mixing preferences may be more prevalent at more localised scales [1, 6]. It is reasonable to suggest then that the gender mixing preferences may vary between groups or organisations, each of which may be represented as a distinct social network. However, making comparisons across networks can be non-trivial and is a problem that has thus far received little attention [3]. Here we address this issue by developing a method for making meaningful comparisons of mixing preferences across networks.

Quantifying the level of assortativity or disassortativity (the preference of linking to nodes with different attributes) can shed light on the organisation of complex networks. It is common practice to measure the level of assortativity according to the Newman's assortativity coefficient [5], the network analogy of Pearson's correlations for attributes across edges. Accordingly, the assortativity coefficient is normalised to lie in the range $r \in [-1, 1]$, where $r = 1$ indicates perfect assortativity with only links between nodes of the same type and $r = -1$ indicates perfect disassortativity in which links only connect nodes of different types. However, when applied to categorical attributes, such as gender, we find that properties of the network imposes more restrictive bounds on the possible range of assortativity values such that the extremal values of 1 or -1 are no longer attainable [2]. Differences in the relative group size are an important factor in this effect. This presents a problem when comparing assortativity across networks as changes in assortativity are confounded with differences in the network structure.

The difficulty associated with comparing gender assortativity across networks is exacerbated when the group sizes are imbalanced. This is of particular concern because the level of assortativity has recently been shown to have an effect on the visibility of a minority group in a network [4]. In science, where women are under-represented, it becomes difficult to compare different organisations (represented by different networks) or to evaluate the impact of policy changes when the groups sizes and connectivity are changing.

Here we propose a solution (details omitted for space reasons) based on normalising the marginal link distribution incident on each gender group. Figure 1 displays a comparison between Newman's assortativity and our proposed normalised variant.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

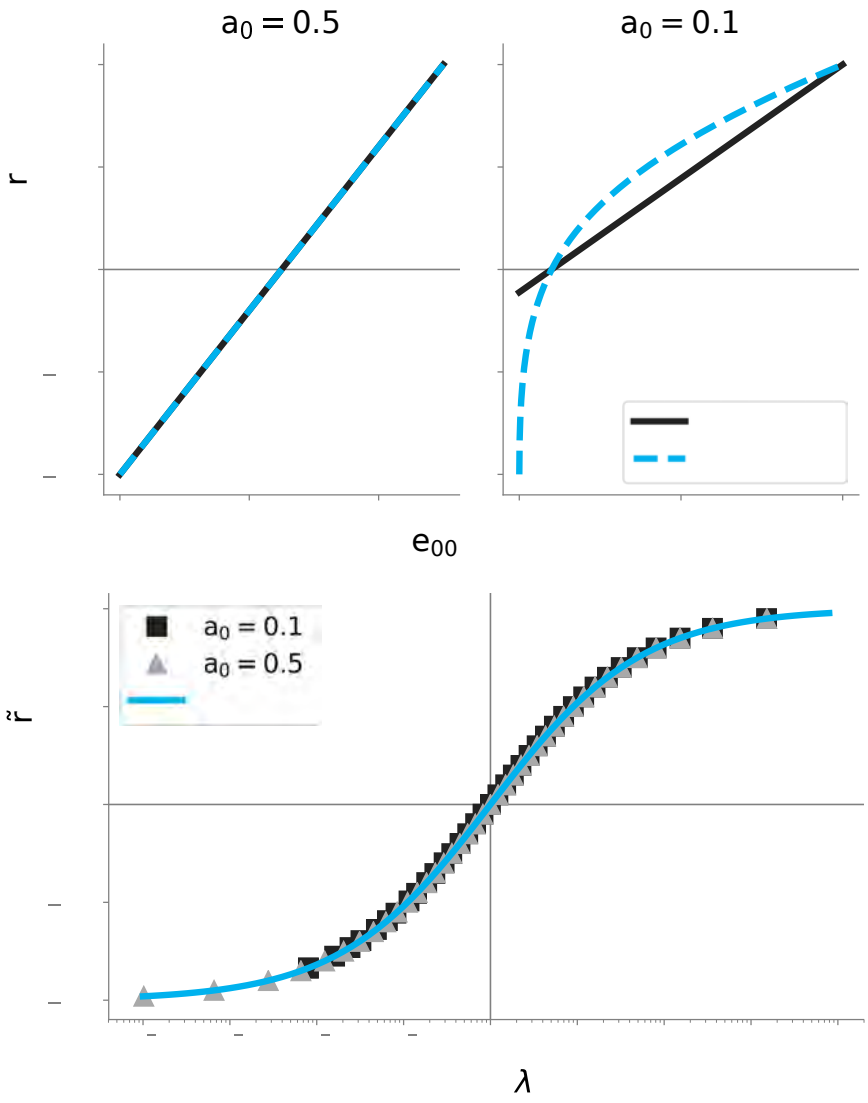


Fig. 1. (A) Two examples of normalized assortativity. When the proportion of edges incident on each group is balanced ($a_0 = a_1 = 0.5$) the original Newman's assortativity and normalised assortativity coincide. When they are imbalanced (e.g. $a_0 = 0.1$), the normalised assortativity is no longer linear, but instead a smooth function that permits the full range of assortativity ($\tilde{r} \in [-1, 1]$) and preserves the same definition of random mixing ($r = 0$). (B) The normalized assortativity as a function of the ratio of within group edges to across group edges. We see a consistent definition of assortativity as a function of the ratio, irrespective of the group sizes.

Consequently we are able to capture and qualitatively evaluate the distribution of mixing patterns across different networks in a population (see Fig. 2).

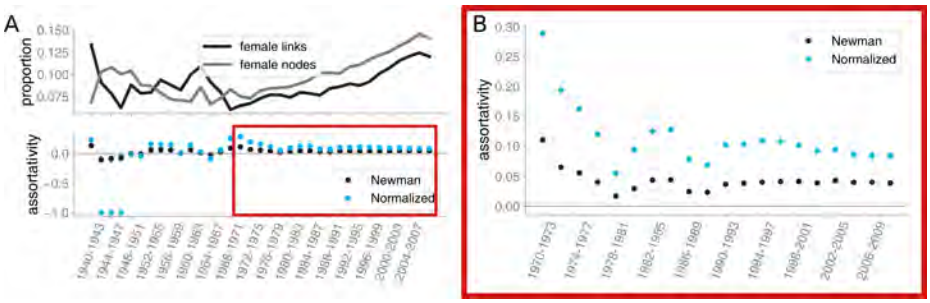


Fig. 2. The gender assortativity of the APS collaboration network over time.

References

1. Altenburger, K.M., Ugander, J.: Monophily in social networks introduces similarity among friends-of-friends. Nature human behaviour 2(4), 284 (2018)
2. Cinelli, M., Peel, L., Iovanella, A., Delvenne, J.C.: Network constraints on the mixing patterns of binary node metadata. arXiv preprint arXiv:1908.04588 (2019)
3. Jacobs, A.Z.: Assembly in populations of social networks. arXiv preprint arXiv:1811.01452 (2018)
4. Lee, E., Karimi, F., Wagner, C., Jo, H.H., Strohmaier, M., Galesic, M.: Homophily and minority-group size explain perception biases in social networks. Nature human behaviour pp. 1–10 (2019)
5. Newman, M.E.: Mixing patterns in networks. Physical Review E 67(2), 026126 (2003)
6. Peel, L., Delvenne, J.C., Lambiotte, R.: Multiscale mixing patterns in networks. Proceedings of the National Academy of Sciences 115(16), 4057–4062 (2018)
7. Ugander, J., Karrer, B., Backstrom, L., Marlow, C.: The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503 (2011)

Dimensions of Social Exchange

Luca Maria Aiello

Nokia Bell Labs, Cambridge, UK
lajello@gmail.com,
WWW home page: http://www.lajello.com

Introduction

Social relationships are among the most important things in our life. They determine and relate to who we marry, where we work, and what we make. They take center stage in our digital lives too. Social-networking sites are made of relationships, and the act of maintaining them results into bridging and bonding forms of social capital and, ultimately, into well-being. Researchers have tried to capture the nuances of relationships by measuring them in terms of tie strength. Yet not all ties of the same strength are created equal. Many social factors are too intertwined to consider tie strength a complete or even a distinctive characterization of a relationship. In this study, we set out to 1) study how people perceive the richness of their relationships beyond tie strength by identifying the main dimensions that define social interactions and 2) develop machine learning tools that are able to infer those interaction types from conversational text.

The 10 dimensions of social exchange

We reviewed the relevant literature in sociology and social psychology and obtained eight tentative dimensions along which relationships could be classified. Independently, we asked 100 crowd-sourcing users to describe their relationships with words and obtained 1,352 terms, 220 of which were unique. We then asked another set of 100 crowd-sourcing users to validate each of these 220 terms through a structured survey. As a result of the crowdsourcing, each word has been characterized by a 100-dimensional rating vector that allowed us to compute the relatedness of words and extract cohesive groups of terms. The groups we found overlap to a large extent with the eight dimensions we found in the social pyschology literature and add two new dimensions. The final list [5] consists of 10 dimensions: *similarity [10], social support [8], trust [14], romance [3], identity [12], respect [7], knowledge [8], power [2], fun [11], and conflict*.

Descriptive and predictive power of social dimensions

To show how this nuanced classification can be used to enchance network science applications, we run a study using a dataset [1] of textual conversations between linked individuals in an online social network. Each dimension is associated to a set of terms from the crowdsourcing; therefore, for each social tie, we were able to match the terms that reflect each of the 10 dimensions, with the words occurring in the conversation. We label each edge with the dimension having the highest number of matching words. We

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

selected 100k connected pairs (positives) and 100k disconnected ones at 2 hops away (negatives) to run a link prediction experiment in two scenarios. In the first, we predict the presence of a link from *A* to *B* based on their common neighbors count *CN*. In the latter, we use a feature vector whose entries count the number of common neighbors who are connected to *A* with a link of a given type (e.g., "support"). In a supervised learning setting with 10-fold cross validation, the latter scenario brings an improvement of 9% in AUC compared to pure CN. Decomposing the tie strength (number of common friends) into its components improves our ability to predict the network structure. The improvement is significant; in link recommendation a +1% in AUC, on a large scale, leads to a large increase in the number of links created.

In addition, when analyzing the sub-graph induced by links of a given type, we find that network properties vary as one would expect from social psychology theories. For example, the network of knowledge exchange tends to be assortative whereas the network of respect is disassortative (people who have high "reputation" are given status mostly by less-respected members of the same community).

Learning the 10 dimensions from text

Finally, to go beyond simple word-matching strategies, we trained a classifier that is able to label conversational text according to the 10 sociological dimensions we identified. To perform the training in a supervised fashion, we collected labeled data using two approaches.

First, we collected 10k comments from reddit.com extracted at random from all the reddit comments posted in 2017 and trained Mechanichal Turk workers to label these comments with any of the 10 dimensions. Each comment was labeled by at least three workers and we considered positive examples those labeled with the same dimension by at least two workers.

In the second approach, we have developed an online platform (www.tinghy.org) where users login to play through Twitter, their timeline data is accessed and they are sequentially presented with 10 of their actual friends. For each friend, they rate the extent to which that relationship is described by our 10 blocks. The user interface is "gamified" so that the experience is fun and rewarding. This platform allowed us to collect conversational data (i.e., mentions) that are implicitly labeled with the 10 fundamental dimensions. So far, we collected data from 500+ users.

Using the data collected and a variety of classifiers (XGBoost [4] trained on a number of NLP features; LSTM [9] and BERT [6] trained of word and sentence embeddings [13]), we achieved very encouraging prediction results in terms of AUC (up to 0.85) when training independent binary classifiers for each individual dimension . In the future, we plan to make the crowdsourcing data and the prediction model available to the community to enable network scientists to study the nuances of social exchange in conversation networks.

References

1. L. M. Aiello, A. Barrat, C. Cattuto, G. Ruffo, and R. Schifanella. Link creation and profile alignment in the anobii social network. In *2010 IEEE Second International Conference on*

Social Computing, pages 249–256. IEEE, 2010.
2. P. M. Blau. *Exchange and power in social life*. Transaction Publishers, 1964.
3. D. M. Buss. *The Evolution of Desire: Strategies of Human Mating*. Basic books, 2003.
4. T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.
5. S. Deri, J. Rappaz, L. M. Aiello, and D. Quercia. Coloring in the links: Capturing social ties as they are perceived. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):43:1–43:18, Nov. 2018.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
7. R. M. Emerson. Social Exchange Theory. *Annual review of sociology*, 2(1):335–362, 1976.
8. S. T. Fiske, A. J. Cuddy, and P. Glick. Universal Dimensions of Social Cognition: Warmth and Competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.
9. F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.
10. M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual review of sociology*, 27(1), 2001.
11. L. Spencer and R. Pahl. *Rethinking Friendship: Hidden Solidarities Today*. Princeton University Press, 2006.
12. H. Tajfel. *Social Identity and Intergroup Relations*. Cambridge University Press, 2010.
13. J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.
14. A. Zaheer, B. McEvily, and V. Perrone. Does Trust Matter? Exploring the Effects of Interorganizational and Interpersonal Trust on Performance. *Organization science*, 9(2):141,159, 1998.

The closed loop between opinion formation and personalised recommendations

Wilbert Samuel Rossi[1], Jan Willem Polderman[2], and Paolo Frasca[3]

[1] Department of Sciences, University College Groningen, University of Groningen, 9718 BG Groningen, The Netherlands, w.s.rossi@rug.nl
[2] Department of Applied Mathematics, University of Twente, 7500 AE Enschede, The Netherlands, j.w.polderman@utwente.nl
[3] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, GIPSA-lab, 38000 Grenoble, France, paolo.frasca@gipsa-lab.fr

Summary. In online platforms, recommender systems are responsible for directing users to relevant contents. In order to enhance the users' engagement, recommenders adapt their output to the reactions of the users, who are in turn affected by the recommended contents. The aim of this work is to make explicit the feedback loop between the evolution of the user's opinion and the personalised recommendation of contents. While our work – described fully in [8] – does not consider a social network for the sake of analytical tractability, similar ideas can be applied to more complex situations where recommender systems mediate social interactions.

1 Introduction

Recommendation systems are ubiquitous in all kinds of web services, such as search engines, social networking service, e-commerce platforms. Their purpose is sieving the information available to them and provide the user with the most relevant items. As online activities become more prominent in the lives of the people, questions are asked about the effects (if any) of recommendation systems on the online and offline behaviors of the users. Our investigation specifically questions the role of personalization.

The issue of personalization is specially perceived as relevant when it comes to the access to news. While on one side personalization enhances user experience, on the other side political activists and scholars have raised concerns that excessive personalisation narrows down the positions available to users about specific issues, effectively enclosing users into so-called "filter bubbles" that favour the emergence of opinion polarisation and radicalisation [5, 6]. Even though this concern has been downplayed by subsequent research [1], it is clear that personalization has at least the potential to reinforce the user's idiosyncrasies and biases, like the *confirmation bias*.

We propose a tractable mathematical model of the interplay between a user and a learning system that provides her with personalized recommendations, and quantify the reciprocal reinforcement of confirmation bias and personalized curation. Our work – described fully in [8] – is related to several recent papers that have tried to incorporate some models of online platforms in models of opinion dynamics [2, 3, 7]. For the sake of analytical tractability, our model neglects the network effects induced by the interaction of multiple users connected via social ties: we indeed believe that our model is a step toward the investigation of such more complex scenario.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Model

We model the opinion formation process of a user that reads news from a news aggregator that provides personalized recommendations, see Figure 1. We restrict ourselves to news that bear implications for one specific issue, say, highlighting the benefits/drawbacks of immigration. News articles are characterized by a (binary) attribute that defines their positive or negative *position* $\mathrm{p}_{\mathrm{art}}(t)$ on the given issue. The *opinion* of the user $\mathrm{o}_{\mathrm{usr}}(t)$ evolves as an affine system that integrates the received news (actually, their positions) along time. Owing to the confirmation bias, i.e. the unintentional tendency to acquire and process evidence that confirms one's beliefs, news items are clicked upon $\mathrm{clk}(t)$ with a probability that is larger when their position is closer to the current user opinion. The recommender system has the objective of improving the engagement of the users, measured as the number of clicks. In order to achieve this purpose, the recommender tracks the number of times that a specific position has been recommended ($\mathrm{r}_+(t)$, $\mathrm{r}_-(t)$) and clicked upon ($\mathrm{a}_+(t)$, $\mathrm{a}_-(t)$). The recommender follows a randomized strategy that, based on these counts, balances "exploration", that is, identifying which position is more appreciated by the user, with "exploitation", that is, providing the user with news that are most likely to be clicked on. Hence, the recommender systems responds to user behavior.

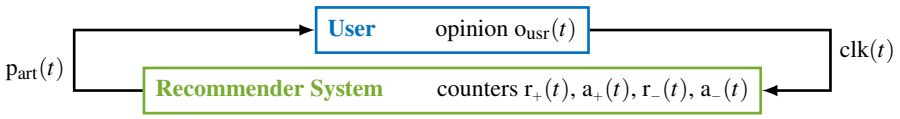
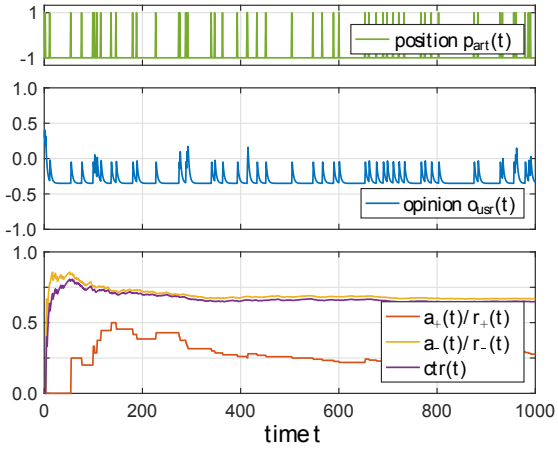


Fig. 1. The closed loop between the user and the news aggregator. The diagram includes the variables exchanged by the two interacting dynamical systems, and their internal state variables.

3 Results and Conclusion

We observe that typical trajectories of the dynamical model are characterized by a definite majority of either positive or negative recommendations, see e.g. Figure 2. Such observation supports the analysis of the expected dynamics *conditioned upon a given majority*: these conditional expectations can be derived in closed-form and turn out to describe the stochastic dynamics very accurately. Statistically, we observe that recommendations produce a significant polarizing effect on the opinions and that this effect is closely entangled with their effectiveness in terms of increasing the click-through rate. Hence, our analysis suggests that mitigating the impact of the recommender system on the opinions has a price in terms of the achievable click-through rate.

While we believe that our model is relevant to the heating debate on the impact of machine learning on our societies, we are well aware of its limitations. Indeed, our model describes the behavior of a single user, but real recommender systems deal with large numbers of users that can have social ties and shared interests. Our recommender system is not allowed to exploit neither of them while real recommender do [4]. Moreover, recommendations are the only drive to the opinion dynamics in our model. Instead,



page 6 - SIM SECTION - Dynamics - reference parameters, down

Fig. 2. A simulation of our model where the majority of recommended articles has negative position (top plot). Consequently, the user opinion becomes negative regardless of its initial positive prejudice (middle plot). Subject to the confirmation bias, this user favours articles with negative position. The recommender recognizes that by computing the acceptance rates of the different positions (lower plot) and in this case continues to recommend mostly article with negative position, to exploit the user preference and maximize the clickthrough rate $\mathrm{ctr}(t) := (\mathrm{a}_+(t) + \mathrm{a}_-(t))/t$.

opinion dynamics are also driven by a network of social interactions (both directly and through the recommender system), creating a complex entanglement of effects.

On this matter, experimental studies on Facebook have reported that ideological contents are primarily filtered by user's social connections rather than by the curation algorithms, suggesting that user preferences may have stronger impact than algorithmic personalisation [1]. A future model that includes both social and recommendation effects, like in the recent paper [7], could shed more light on this issue.

References

1. E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.
2. P. Bolzern, P. Colaneri, and G. De Nicolao. Opinion influence and evolution in social networks: A Markovian agents model. *Automatica*, 100:219 – 230, 2019.
3. P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.
4. D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.
5. D. Lazer. The rise of the social algorithm. *Science*, 348(6239):1090–1091, 2015.
6. E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.
7. N. Perra and L .E.C. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific reports*, 9(1):7261, 2019.
8. W. S. Rossi, J. W. Polderman, and P. Frasca. The closed loop between opinion formation and personalised recommendations. *arXiv preprint arXiv:1809.04644*, 2018.

Network based Modelling and Analysis of Film Performance in the Indian Film Industry

Samrat Gupta[1] and Amit Anand Tiwari[2]

[1]Indian Institute of Management, Ahmedabad, Gujarat 380015, India
samratg@iima.ac.in,
[2]Indian Institute of Management, Rohtak, Haryana 124010, India
amit.tiwari@iimrohtak.ac.in

1 Introduction

Several complex tasks require some form of coordinated collective action to produce non-rival goods such as music, cinema, national defense etc. The relationships rooted in collective action for production of these non-rival goods are based on emotional and cognitive ties [13]. One such example is making of films wherein film performance is not only dependent on individual creative talent but also on direct or indirect relationships among the film professionals (i.e. their network structure). This network structure of professionals working on a film is revealed through their prior film collaborations. Surprisingly, India produces more films and accounts for the largest number of cinema goers compared to any other country in the world. However, in terms of performance which can be assessed either on the basis of the quality of content as assessed by legitimate users or on the basis of its box-office earnings, American (US/Canada) and Chinese film industries are far ahead of India [5]. Intrigued by this observation and in our quest to understand the relational configurations that effect performance of movies released in India, we curate the year-wise network data of professional collaborations in movies during two decades (2000 – 2019) from Internet Movie Database (IMDb) and study the properties and mesoscopic structures within these networks.

We pose following research questions so as to emphasize the focus of our work.

1) What are the temporal network characteristics of collaboration network of movies released in India?
2) On what basis are communities organized in film collaboration network and is there any correspondence between community structure and film performance? To this end, we propose a new fuzzy-rough set based community detection algorithm for weighted networks.
3) What is the relative performance of proposed approach as compared to state-of-the-art methods for community detected in weighted networks?

Addressing these questions, we expand the past research which has mainly focused on box office performance as a function of variables related to value chain of movie such as genre, screen count, advertising etc. [2,3,5].

2 Proposed Methodology

First we form weighted networks for each year starting from 2000 to 2019 as there was a marked improvement in the quality of production of Indian cinema in 2000 due to

Table 1: Summary of Twenty Collaboration Networks

Movie Released (Year Window)	Movies in Network	# of Movie Professionals	# of Links	# of Communities Detected
1997-2000	1694	5193	44648	721
1998-2001	1828	5640	46467	814
1999-2002	1957	6135	49313	784
2000-2003	2062	6600	52735	939
2001-2004	2156	7118	57482	1080
2002-2005	2213	7592	62118	982
2003-2006	2171	7980	65247	1131
2004-2007	2142	8371	67204	1186
2005-2008	2145	8937	69947	1291
2006-2009	2148	9564	72947	1245
2007-2010	2367	10768	80222	1608
2008-2011	2569	11845	86253	1792
2009-2012	2765	12960	92491	1937
2010-2013	3092	14537	102119	2197
2011-2014	3367	15853	108780	2083
2012-2015	3730	17597	119384	2641
2013-2016	4117	19547	131135	2957
2014-2017	4292	20753	136119	3080
2015-2018	3852	19429	125740	2876
2016-2019	3099	16766	104185	2467

technological advancements in cinematography, story line, special effects and animation. Given that collaboration in film industry is characterised by rapid construction and disintegration on project by project basis [10], we control for relationship decay using a three year moving window [1]. As shown in Table 1, for a given year (say 2004), its collaboration network consists of all the film collaborations that took place during last three years and that year (2001-2004). We use the resulting twenty time-varying weighted networks to compute network properties and reveal community structure.

To account for bias in edge weights due to popularity of film professionals, we follow a two-step normalization process [15]. First, we normalize an edge weight between two professionals v_i and v_j by setting the weight as

$$w'_{ij} = \frac{w_{ij}}{m_i * m_j} \tag{1}$$

where w_{ij} is the total number of movies in which v_i and v_j have collaborated, m_i and m_j are the total number of movies on which v_i and v_j have worked. In the second step, we normalize all w'_{ij} by dividing each edge weight with the maximum edge weight obtained from first step. Thus normalized adjacency matrix of a network is given as:

$$A_{ij} = \begin{cases} w'_{ij} \Big/ max_{\forall(i,j)} \{w'_{ij}\} & \text{if node } v_i \text{ connects to node } v_j \\ 0 \quad \text{otherwise} \end{cases} \tag{2}$$

An example of edge normalization using our two-step normalization process is shown in Figure 1. Once, normalized network is obtained, weighted neighborhood subset (WNS) of each node in the network is formed. Subsequently, constrained connectedness upper approximation subsets based on a concept related to rough set theory [11]

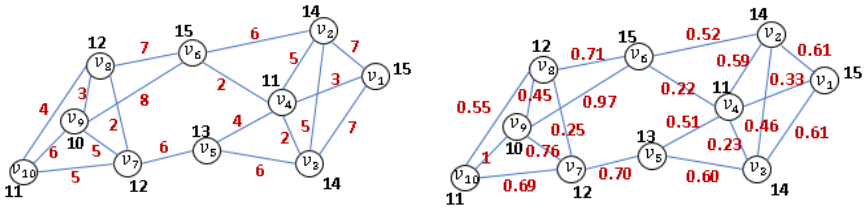


Figure 1. (a) Weighted Toy Network (b) Toy Network with Normalized Weights

are computed by iterating until convergence. The concept of weighted relative connectedness (WRC) (as shown in Eq. 3) is used to constrain and merge the sets during each iteration. This notion of WRC is used to compute similarity between every pair of nodes and filter out the nodes for which WRC $\leq$ δ in each iteration (where δ is a user-defined threshold and $\delta = 1$ for toy network).

$$WRC(v_i, v_j) = \frac{|WNS(v_i) \cap WNS(v_j)|}{\min(|WNS(v_i) - WNS(v_j)|, |WNS(v_j) - WNS(v_i)|)} \quad (3)$$

For better understanding, we illustrate the computation of weighted relative connectedness between nodes v_7 and v_9 of a toy network shown in Figure 1. The weighted neighborhood subsets of v_7 and v_9 can be denoted as $WNS(v_7) = \{(5,0.70), (8,0.25), (9,0.76), (10,0.69)\}$ and $WNS(v_9) = \{(6,0.97), (7,0.76), (8,0.45), (10,1)\}$ respectively. Now, using the concepts of fuzzy set theory [14], weighted relative connectedness between v_7 and v_9 can be calculated as follows:

$$|WNS(v_7) \cap WNS(v_9)| = |\{(8,0.25), (10,0.69)\}| = 0.94$$
$$|WNS(v_7) - WNS(v_9)| = |WNS(v_7) \cap WNS(v_9)^c| = |(5,0.70)(8,0.25)(9,0.76)| = 1.71$$
$$|WNS(v_9) - WNS(v_7)| = |WNS(v_9) \cap WNS(v_7)^c| = |(6,0.97)(7,0.76)(8,0.45), (10,0.31)| = 2.49$$
$$WRC(v_7, v_9) = 0.94/(\min(1.71,2.49)) = 0.94/1.71 = 0.549$$

The synergistic use of WRC and upper approximation identifies meaningful communities in a weighted network. As expected, two overlapping communities viz. (1,2,3,4,5,6) and (6,7,8,9,10) were identified in toy network by the proposed method. To further evaluate the proposed method, we conducted experiments on benchmark weighted networks viz. Karate club network, SFI collaboration network, Les Miserables, C. elegans neural network and US Air Transportation network [4,6,12]. The detected community structure in these networks is consistent and coherent with the respective ground truth structure.

We also study structural properties of these weighted networks such as power law degree distribution, local and global clustering coefficients, betweenness and eigenvector centralities, average path length, structural holes, embeddedness, community structure and rich club effect [8]. For studying these weighted network properties, we use more sophisticated measures as compared to traditional measures for unweighted networks [7,9]. The examination of year-wise distribution of movies in terms of genre, average and variance of year-wise movie ratings also reveals interesting observations.

3 Research Findings and Implications

This research has several findings that can aid producers and movie studios in producing commercially and/or artistically viable content at the box-office. The findings suggest that group performance surfaces across structural holes and network closure. Centrality analysis reveals that lesser popular actors who appear quite frequently for negative or comic roles in Indian movies have higher eigenvector centrality. Since, the eigenvector centrality connects focal individual to many others (directly and indirectly), without being resource intensive in managing focal individual's network, this finding implies that if one is a good character actor, then that person can work with stars, who themselves may not work with each other. Further investigation reveals that collaboration networks of movies released in India do not follow weighted rich club effect (Figure 2). This finding indicates that prominent movie professionals in India do not share their strongest ties with other prominent professionals rather with less prominent professionals. The proposed community detection approach identifies low-budget, high budget, low performing and high performing movie collaborations in the Indian film industry. Experiments and comparative analysis with state-of-the-art algorithms conducted on real weighted networks show that proposed approach provides significant improvements in identifying communities within weighted networks. This research has several managerial implications such as providing guidance to film makers in maximizing revenues through strategic assembly of movie team, predicting the future collaboration patterns of film professionals and deriving meaningful insights about the controversial issues such as nepotism in Indian film industry. Further, this research work driven by real-world data has instructional value to similar research areas such as financial contagion in banking system and brand advertising on social media platforms where business networks may be studied.

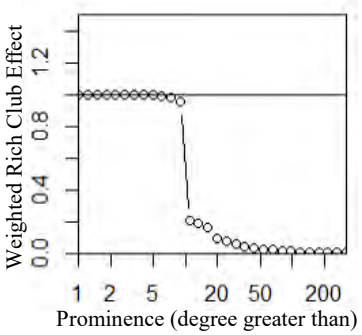


Figure 2. Absence of weighted rich club effect for the collaboration network of movies during 2014-2017 (Similar effects were observed in collaboration networks of all other year windows)

References

1. Cattani, G., Ferriani, S.: A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the Hollywood film industry. Organization Science 19(6), 824-844, (2008)

2. Delen, D., Sharda, R.: Predicting the financial success of hollywood movies using an information fusion approach. Industrial Engineering Journal 21(1), 30–37, (2010)
3. Eliashberg, J., Elberse, A., Leenders, M. A.: The motion picture industry: Critical issues in practice, current research, and new research directions. Marketing Science 25(6), 638-661, (2006)
4. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821-7826, (2002)
5. Gupta, S., Kumar, S., Kumar, P.: Evaluating the Predictive Power of an Ensemble Model for Economic Success of Indian Movies. Journal of Prediction Markets 10(1), (2016)
6. Knuth, D.E.: The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley Reading, 1993
7. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32(3), 245-251, (2010)
8. Opsahl, T., Colizza, V., Panzarasa, P., Ramasco, J.J.: Prominence and control: the weighted rich-club effect. Physical Review Letters 101(16), (2008)
9. Opsahl, T., Panzarasa, P.: Clustering in weighted networks. Social Networks 31(2), 155-163, (2009)
10. Packard, G., Aribarg, A., Eliashberg, J., Foutz, N.Z.: The role of network embeddedness in film success. International Journal of Research in Marketing 33(2), 328-342, (2016)
11. Pawlak, Z.: Rough Sets. International Journal of Computer & Information Sciences 11(5), 341-356, (1982)
12. Radicchi, F., Ramasco, J.J., Fortunato, S.: Information filtering in complex weighted networks. Physical Review E, 83(4), (2011)
13. Weick, K.E.: The collapse of sense making in organizations: The Mann Gulch disaster. *Administrative Science Quarterly*, 628-652, (1993)
14. Zadeh, L. A.: Information and Control. Fuzzy sets 8(3), 338-353, (1965)
15. Zhang, K., Bhattacharyya, S., Ram, S.: Large-Scale Network Analysis for Online Social Brand Advertising. MIS Quarterly 40(4), (2016)

Selective Exposure shapes the Facebook News Diet

Matteo Cinelli[1], Emanuele Brugnoli[1] Ana Lucia Schmidt[2], Fabiana Zollo[2], Walter Quattrociocchi[2], and Antonio Scala[1,3]

[1] Applico Lab, CNR-ISC, Rome, Italy,
matteo.cinelli@roma1.infn.it,
[2] Università di Venezia "Ca' Foscari", Venezia, Italy
[3] LIMS, the London Institute for Mathematical Sciences, London, UK

The social brain hypothesis approximates the total number of social relationships we are able to maintain at 150 [1]. Such a theoretical cognitive limitation emerges in several other contexts from the patterns of human mobility to the way we communicate. Furthermore, the uptake of social media has radically changed the way we consume content online. Indeed, the way we consume information and the cognitive limits and algorithmic mechanisms underpinning them have a bearing on foundational issues concerning our news consumption patterns. Recent studies targeting Facebook [2] have shown that content consumption is dominated by selective exposure – i.e. the tendency of users to ignore dissenting information and to interact with information adhering to their preferred narrative –

and that individual choices more than algorithms [3] also characterise the consumption patterns of users and their friends [4].

In such a vein, we perform a thorough quantitative analysis to characterise users' attention dynamics on news outlets on Facebook. In particular, we study how 14 million Facebook users distribute their activity among 50000 posts, clustered by topics, produced by 583 pages listed by the Europe Media Monitor over a six-year time span.

We find that users, independently of their activity and of the time they spend online, show a tendency to interact with a very limited number of news outlets. To test the presence of selective exposure, for which evidence emerges from users focusing their attention on a set of preferred news sources (as shown in the top panels of Figure 1), we analyse how homogeneously users distribute their activity across pages and topics. More precisely, the concentration of the distribution of likes towards a certain page or topic signals the presence of selective exposure, while the heterogeneity of such a distribution determines the strength of selective exposure. Such heterogeneity in the distribution of users' likes is quantified by means of the Gini index [5], a classic example of a synthetic indicator used for measuring inequality of social and economic conditions, that we renormalise for being applied to sparse data [6].

We find that highly engaged users tend to concentrate their activity on few pages while being less selective of the topics presented by the pages. In general, we observe that selective exposure increases in strength when the activity of users (i.e. the number of likes) grows but is not affected by users' lifetime (i.e. the time span between the first and the last like).

Our results suggest that the tendency of users to limit their attention to a smaller number of news sources might be one of the factors behind the emergence of echo chambers online. Such an outcome still underlines the tendency of users towards seg-

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

regation, partly because of their attitude and cognitive limits, and partly because of the features of the social media in which they operate.

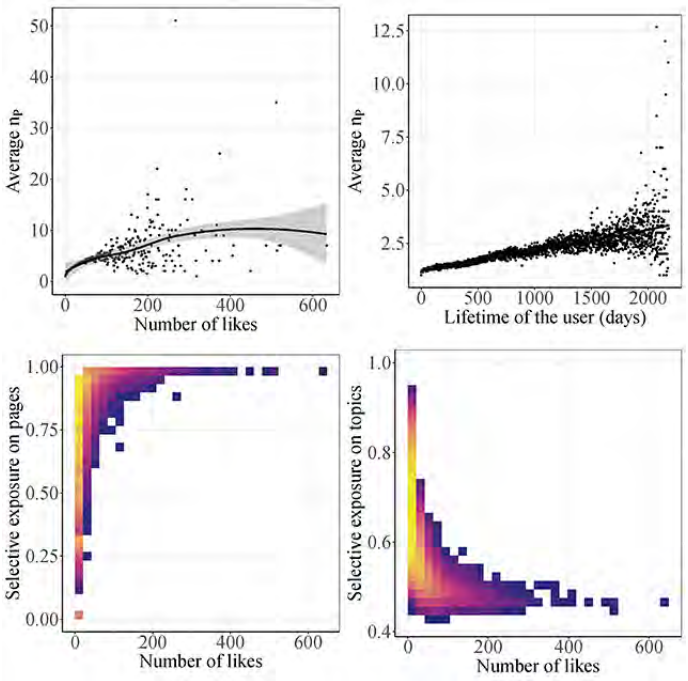


Fig. 1. Top-left panel: relationship between the average number of pages that received likes by users with respect to their activity (quantified by the number of likes). We observe that the average number of pages reaches a plateau of ~ 10 pages for users with an activity of more than ~ 300 likes. Top-right panel: relationship between the average number of pages that received likes by users with respect to their lifetime (quantified by the time between the first and the last like). We observe that the average number of pages grows slowly and reaches a value of ~ 3 pages for most lifelong users. Bottom-right panel: the distribution of selective exposure to pages with respect to users' activity shows that increasing activity levels correspond to higher selective exposure, i.e. users concentrate on fewer pages. Bottom-right panel: the distribution of selective exposure to topics with respect to users activity shows that increasing activity levels correspond to lower selective exposure, i.e. users concentrate on a higher number of topics. Topics are obtained by processing posts using a state-of-the-art topic modeling algorithm.

References

1. Dunbar RI: Social cognition on the Internet: testing constraints on social network size. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012;367(1599):2192–2201.
2. Zollo F, Bessi A, Del Vicario M, Scala A, Caldarelli G, Shekhtman L, et al. Debunking in a world of tribes. PloS one. 2017;12(7):e0181821.

3. Bakshy E, Messing S, Adamic LA. Exposure to ideologically diverse news and opinion on Facebook. Science. 2015;348(6239):1130–1132.
4. Bessi A, Petroni F, Del Vicario M, Zollo F, Anagnostopoulos A, Scala A, et al. Homophily and polarization in the age of misinformation. The European Physical Journal Special Topics. 2016;225(10):2047–2059.
5. Gini C. Measurement of inequality of incomes. The Economic Journal. 1921;31(121):124–126.
6. Cinelli, M., Brugnoli, E., Schmidt, A. L., Zollo, F., Quattrociocchi, W., & Scala, A. Selective exposure shapes the facebook news diet. 2019; arXiv preprint arXiv:1903.00699.

Competing local and global interactions in social dynamics: how important is the friendship network?

Arkadiusz Jędrzejewski, Bartłomiej Nowak, Angelika Abramiuk, and Katarzyna Sznajd-Weron

Department of Theoretical Physics, Wrocław University of Science and Technology, Wrocław, Poland
arkadiusz.jedrzejewski@pwr.edu.pl

1 Introduction

As noted by Kardar and Kaufman: *"The study of competing short-range and long-range interactions is relevant to a variety of problems in statistical mechanics"*. Indeed, one can easily indicate a number of natural processes in which elements interact both locally and globally [1]. Such competing interactions are frequently responsible for the universality of many self-organized patterns observed in condensed matter physics [2, 3]. However, the mutual existence of forces with different length-scales is not only limited to physical or biological systems. In fact, more and more empirical studies are pointing out that the overall social influence results from such a composition of local and global interactions [4–6]. In the era of omnipresent mass media and online social networking, people's interactions are certainly no longer restricted to physical contacts. Their range, in fact, extends easily even beyond geographical borders. This rises a justified question about the significance of these interactions in shaping trends and opinions. Do such forces lead to characteristic macroscopic patterns as their counterparts in condensed matter physics? Can we observe some universal features of social systems with competing social influences? Finally, what is the impact of a social structure in all of this?

Our research builds upon a recent correlation study on social influence in online movie ratings [5]. Having analyzed tendencies among reviewers to conform to already existing comments, the authors reached a conclusion that opinions expressed by friends and strangers cause different social responses. It turned out that those shared by the friends only led to conformity in issued reviews, whereas those of strangers might also excite anticonformity depending on the movie popularity. These findings suggest that some types of social responses may be associated with specific interaction lengths. Concerning a friendship network in this particular study, local interactions with nearest neighbors manifested only conforming nature, whereas those global ones with strangers also displayed anticonforming properties.

Our work is directly inspired by this observation. We have picked one of the prime models in the field that already incorporates these two types of social responses, and we have checked how different constraints on the interaction ranges impact its behavior in a stationary state. We have examined the model on different complex networks generated by Watts and Strogatz's algorithm [7], hoping to also determine the role of the social

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

structure in such systems. Monte Carlo simulations are backed up with mean-field and pair approximations.

2 Model description

This study focuses entirely on the q-voter model with anticonformity and conformity introduced in Ref. [8]. In the original model, interactions occur exclusively between voters that are direct neighbors. In the friendship network, it translates to forces between friends. We call such interactions local. In the current study, we also consider global interactions. These are not limited by the network structure, and they can extend throughout the system, reaching also strangers. Although the empirical study suggests which of the social interactions is long-range, we can imagine that it is the social context that dictates the range of forces. Therefore, we compare four q-voter models with different combinations of local and global sources of conformity and anticonformity. In all cases, social influence originates from a unanimous group of q distinctive voters. However, depending on a considered interaction range, members of this group are randomly selected at the local or global level.

3 Results

The parameters of studied systems have been chosen to accord with psychological theories of social responses, and they reflect properties of real structures. In systems with

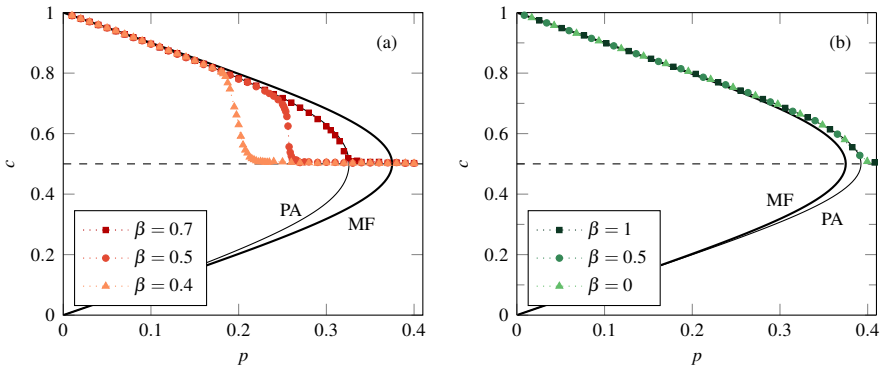


Fig. 1. Phase diagrams for dynamics with (a) global anticonformity and local conformity and (b) local anticonformity and global conformity on Watts-Strogatz networks with $N = 28160$ nodes, the average node degree $\langle k \rangle = 50$, and different values of rewiring probability β. The group of influence consists of $q = 4$ members for all cases. The concentration of voters with one of two possible opinions is denoted by c, whereas the control parameter, which represents the level of anticonformity in the system, by p. Solid thick and thin lines illustrate mean-field (MA) and pair (PA) approximations, respectively. Marks correspond to Monte Carlo simulations.

global anticonformity and local conformity, the majority opinion is the most sensitive to structural changes in the friendship network (see Fig. 1a), and its formation is possible on the smallest interval in the parameter space. A system that exhibits such interactions is reported in the cited study on movie ratings. In contrast, combining local anticonformity with global conformity makes the majority opinion more resistant to structural changes (see Fig. 1b). In fact, the influence of the network structure on the final opinion is negligible for the parameters that characterize many real social systems. In these cases, only the average number of friends in the population impacts the outcome. Although the limiting behavior of all the dynamics is the same, the differences between them are noticeable for the typical values of the average node degree found in real-world structures. Thus, if the models of opinion dynamics intend to properly capture the collective human behavior, it is important to accurately determine the ranges of social interactions since they can completely change the system properties.

Acknowledgments

This work was created as a result of the research projects nos. 2016/21/B/HS6/01256, 2016/23/N/ST2/00729, and 2018/28/T/ST2/00223 financed from the funds of the National Science Center (NCN, Poland). This research was supported in part by PLGrid Infrastructure.

References

1. González-Avella, J.C., Eguíluz, V.M., Cosenza, M.G., Klemm, K., Herrera, J.L., San Miguel, M.: Local versus global interactions in nonequilibrium transitions: A model of social dynamics. Phys. Rev. E 73, 046119 (2006)
2. Stoycheva, A.D., Singer, S.J.: Stripe Melting in a Two-Dimensional System with Competing Interactions. Phys. Rev. Lett. 84, 4657–4660 (2000)
3. Seul, M., Andelman, D.: Domain Shapes and Patterns. The Phenomenology of Modulated Phases. Science 267, 476–483 (1995)
4. Onnela, J.-P. and Reed-Tsochas, F.: Spontaneous emergence of social influence in online systems. Proc. Natl. Acad. Sci. U.S.A. 107, 18375–18380 (2010)
5. Lee, Young-Jin and Hosanagar, Kartik and Tan, Yong: Do I follow my friends or the crowd? Information cascades in online movie ratings. Manag. Sci. 61, 2241–2258 (2015)
6. Pan, X., Hou, L., Liu, K.: Social influence on selection behaviour: Distinguishing local-and global-driven preferential attachment. PLOS ONE 12, e0175761 (2017)
7. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440 (1998)
8. Nyczka, P., Sznajd-Weron, K., Cisło, J.: Phase transitions in the q-voter model with two types of stochastic driving. Phys. Rev. E 86, 011105 (2012)

Mixing dynamics and group imbalance lead to degree inequality in face-to-face interactions

Marcos Oliveira[1], Fariba Karimi[1], Maria Zens[1], Johann Schaibl[1], Mathieu Genois[3], and Markus Strohmaier[2]

[1] GESIS- Leibniz Institute for the Social Sciences, Cologne, Germany,
[2] Dept. for Society, Technology and Human Factors
RWTH Aachen University, Germany
[3] Centre de Physique Theorique, Campus de Luminy, Marseille, France.

1 Introduction

Homophily plays a significant role in shaping social structure and in influencing dynamics on social networks. Recently, researchers have traced a link between homophily and minorities, revealing that homophily accentuates underrepresentation in rankings of social networks with minority groups. In this paper, we study the impact of such dynamics on face-to-face interactions. Precisely, we characterize discrepancies in the interactions of minorities and majorities, and subsequently develop a model to explain them. First, we expose some characteristics of the networks that emerge from face-to-face interactions: degree distribution, strength distribution, and contact duration distribution. In line with previous studies, we find degree inequality emerging as a consequence of social interactions. We argue that besides attractiveness, homophily plays a significant role in these differences. We evaluate attribute assortativity and the connectivity between classes. Finally, we propose a network model of face-to-face interactions based on attractiveness and homophily. We show that the discrepancies in the data can be explained by the addition of homophily in the model.

2 Results

We studied the social networks of schools and conferences that used sociopattern proximity sensors to collect face-to-face interactions[1, 2]. With these data sets, we built the social networks in which a node is a person, and an edge indicates interaction between two people. In these networks, the degree distributions are well behaved around a center tendency. The data also contains the gender information. In all considered cases, there exist less female students than male students.

Degree inequality and mixing in social networks

We first characterized the group connectivity patterns in the social networks. For this, we measured the average degree of each group in the networks (Fig. 2A).

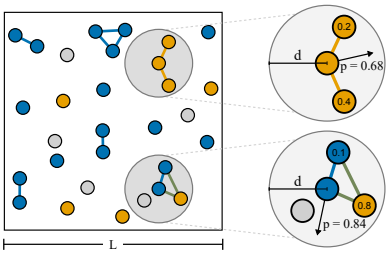


Fig. 1. Schematic description of the attractiveness–mixing model.

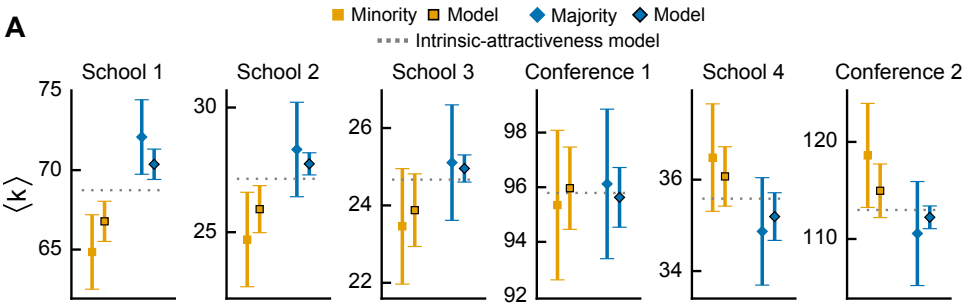


Fig. 2. Degree inequality and mixing in face-to-face social networks. The empirical average degree of the minority and majority is compared with the model

We found a systematic degree inequality among groups. The minorities exhibit lower average degree than the majorities in all classes but School 5, in which the opposite occurs. The previous model of face-to-face interactions in space with intrinsic attractiveness of the individuals fails to explain this observation[3] as it neglects relational attributes in social dynamics.

Here, we present a social network model of physical proximity that incorporates (i) intrinsic attributes of individuals and (ii) relational attributes between groups. We show that these ingredients are sufficient to explain degree inequality observed in social dynamics with minorities. In this model, each individual has an intrinsic attractiveness that is drawn from a uniform distribution. The members of a group share the same mixing pattern, which tunes how individuals interact with others. In general, individuals move across the space depending on their label and the composition of their surroundings (see Fig. 1). While the previous intrinsic-attractiveness model proposes that individuals are more likely to interact with high intrinsically attractive individuals, here we argue that this likelihood also depends on the mixing dynamics between the groups.

In the attractiveness–mixing model, each individual has three attributes: a label $b_i \in [0, B-1]$, where B is the number of groups; an intrinsic attractiveness $\eta_i \in [0, 1]$; and an activation probability $r_i \in [0, 1]$. The mixing patterns in this system are encoded in the $B \times B$ mixing matrix $\mathbf{h}$. Each row of $\mathbf{h}$ can be seen as a probability mass function that weighs the likelihood of group interaction.

In the model, N individuals perform random walks in a two-dimensional $L \times L$ periodic space and move based on the composition of their vicinity. For this, we define $\mathsf{N}_i(t)$ as the set of individuals who are within radius d from the individual i at time t. The individuals move only probabilistically. At each time step t, each individual i moves with probability

$$p_i(t) = 1 - \max_{j \in \mathsf{N}_i(t)} \{\eta_j h_{b_i b_j}\}, \tag{1}$$

In this model, an individual interacts with others depending on their *perceived* attractiveness—as perceived by the group of this individual. Each individual moves with a step of length v along a random direction of angle $\xi \in [0, 2\pi)$. Finally, individuals can be active or inactive; they only move and interact with others if they are active. An inactive individual i becomes active with probability r_i, whereas an active but *isolated* individual i becomes inactive with probability $1 - r_i$. In this study, we assume that the intrinsic attractiveness η_i and the activation probability r_i come from a continuous uniform distribution in $[0, 1]$.

Our results suggest that in order to have more accurate models of social interactions in physical proximity it is crucial to account for mixing patterns between the groups. In addition, we show how these mixing patterns result in surprising degree ranking inequalities for minorities and majorities.

References

1. Julie Fournet and Alain Barrat. Contact patterns among high school students. *PloS one*, 9(9):e107878, 2014.
2. Mathieu Génois, Maria Zens, Clemens Lechner, Beatrice Rammstedt, and Markus Strohmaier. Building connections: How scientists meet each other during a conference. *arXiv preprint arXiv:1901.01182*, 2019.
3. Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. Modeling human dynamics of face-to-face interaction networks. *Physical review letters*, 110(16):168701, 2013.

Coffee Discussion on Twitter: A Sentiment Analysis Taking Network Topology Into Account

Nga Nguyen, Franz-Benjamin Mocnik, and Evangelia G. Drakou

University of Twente, Faculty of Geo-Information Science and Earth Observation,
The Netherlands
n.p.nguyen@utwente.nl,
https://people.utwente.nl/n.p.nguyen

1 Introduction

Climate change awareness plays an important role in behavior change towards a more sustainable future [3]. Social media such as Twitter reflects public awareness as more people are taking on to social platforms to express their opinion [4]. Several efforts have been made to analyze public sentiment based on crowd sourced information. Most of these efforts are in the health care sector and focus on disease awareness and epidemiological spread [1]. There has, however, been less research combining sentiment analysis and social media on the ways information on climate change or sustainability issues spreads. In this paper, we address this gap by relating the awareness about sustainability issues to socially created norms, thereby discussing the roles of hubs and peripherals.

2 Methodology

In order to understand what factors influence consumer sentiment on sustainable coffee, we scraped Twitter as the crowd source for 'psychological wisdom'[6]. We used 5M tweets in the last 10 years from 4,000 users who recently tweeted about coffee in the Netherlands (1M tweets in English).

Each tweet is analyzed for its content and sentiment. First, the tweets are tagged by their content. Tweets containing 'sustainability', 'climate change', etc. are tagged as sustainable tweets; tweets containing 'coffee' are tagged as coffee tweets; within this set of coffee tweets, any mention of 'sustainable', 'organic', 'certified' is tagged as a sustainable coffee tweet. Secondly, we applied sentiment analysis to the tweets based on Syuzhet's sentiment algorithm [7]. Each resulting sentiment score is in the interval $[-1,1]$, with negative sentiment receiving the score of -1 and positive sentiment having a score of 1. The sustainability or sustainable coffee sentiment is then normalized to the user's average sentiment score. Thirdly, a network is created with users as nodes and mentions, including retweets and replies, as edges. Such edges are preferred over a friendship network as they are dynamic and potentially change every month. Fourthly, we run a regression on the sentiment score to understand which factors influence consumer sentiments on sustainable coffee. Then, we divide the data into two sets, hub and peripheral. This discrimination is made based on both network topology and the users' tweeting patterns. The latter of these two factors is important in the context of this paper because it incorporates information about the way Twitter users interact. The

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

same regression as before is run again on both sets in order to compare the behavior between the hubs and peripherals.

According to the Theory of Reasoned Action [2], intention is influenced by attitude and social norms. Here, intention is understood as sentiment/awareness about sustainable coffee, while attitude is sentiment about sustainability issues in general. As neighboring users can have influence on one's sentiment, their sentiment is considered here to be a 'social norm'. To understand the contribution of each of these factors, we ran several regressions: ordinary least square (OLS), random effect and fixed effect. OLS and random effect both perform worse than fixed effect, based on the Hausman test.

We are looking for ways to explain the monthly sentiment score $sc_{u,t}$ about sustainable coffee (with respect to a user u at time t) in terms of the sentiment score $s_{u,t}$ about sustainability (independent of the coffee), as well as both sentiment scores aggregated over the neighborhood $\mathcal{N}(u,t)$. Here Twitter is treated as e-word of mouth [5], which is why the neighborhood has a social influence to the user u and thus potentially opposes Social Norms to u. The final regression thus takes the following format:

$$sc_{u,t} \stackrel{!}{=} \beta_s \cdot s_{u,t} + \bar{\beta}_s \cdot \frac{1}{|\mathcal{N}(u,t)|} \sum_{u' \in \mathcal{N}(u,t)} s_{u',t} + \bar{\beta}_{sc} \cdot \frac{1}{|\mathcal{N}(u,t)|} \sum_{u' \in \mathcal{N}(u,t)} sc_{u',t} + \varepsilon =: \widehat{sc}_{u,t} \quad (1)$$

In addition, we explored whether the current sentiment towards sustainable coffee $sc_{u,t}$ can be explained by the corresponding past sentiment $sc_{u,t-1}$. Equation 1 thus transforms to

$$sc_{u,t} \stackrel{!}{=} \widehat{sc}_{u,t} + \alpha_{sc} \cdot sc_{u,t-1} \quad (2)$$

In both cases, we divide the Twitter users into hubs and peripherals, based on the count of followers, status and favorites. The former is defined as Twitter users who have a high number of followers tweeting frequently and highly-liked content. Hubs have thus the potential to steer the direction of the discussion based on the ability to spread large information over the network. It can be expected to see a difference in information diffusion behavior based on the different network topology.

Similar to the case of the sentiment score, a corresponding regression can be made for the awareness about sustainability. Awareness about sustainability is simply a value of 0 or 1, with 1 mentioning the term and 0 otherwise.

3 Results and Discussion

The fixed effect regression shows that social norms, created by the neighborhoods, have a more significant predictive power toward the sentiment of sustainable coffee than

Table 1. Fixed Effect Regression for Sustainable Coffee Sentiment

Coefficient	Whole network	Hub	Peripheral
β_s	0.005 ± 0.002	0.034 ± 0.013*	0.002 ± 0.003
$\bar{\beta}_s$	0.002 ± 0.002	-0.009 ± 0.000	0.004 ± 0.002
$\bar{\beta}_{sc}$	0.762 ± 0.011***	0.984 ± 0.030***	0.729 ± 0.011***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2. Fixed Effect Regression for Sustainable Coffee Sentiment

Coefficient	Whole network	Hub	Peripheral
β_s	0.000 ± 0.002	0.002 ± 0.004	-0.002 ± 0.003
$\bar{\beta}_s$	0.014 ± 0.001***	-0.001 ± 0.004	0.015 ± 0.003***
$\bar{\beta}_{sc}$	0.554 ± 0.011***	0.999 ± 0.010***	0.532 ± 0.012***
α_{sc}	0.093 ± 0.013***	0.000 ± 0.022	0.096 ± 0.014***

the user's own sentiment of sustainability (Table 1). That is network influence is more important than intrinsic attributes. Further, there is a difference in the behavior of hub and peripherals. Based on the definition given above, we refer to 154 Twitter users as hubs, making up 3.7% of the data set. The hubs correlate almost perfectly with the sentiment of their neighbors, which reflects their role to spread information, whereas for peripherals, this correlation is not as strong.

Based on our analysis incorporating the past sentiment (Table 2), only an insignificant relationship between the lagged term and the sentiment score of tweets containing sustainable coffee could be observed. A possible explanation is the diversity of messages being tweeted or the temporal scale being too granular.

The study shows that the ways a user is able to influence other users' sustainable coffee sentiments depends on the topology of the social network. In fact, we were able to demonstrate that the sentiments about sustainable coffee of users within the neighborhood is far more important than the sentiment of sustainability alone. In case of communication about sustainability issues, it could be traced that hubs are thus very effective (and influential) in affecting other users' sentiments.

We were not able to fully explore the ways past and present sentiments are linked through time. Future research could explore different temporal aggregations, thus assuming another scale of analysis. Further, we would like to include different time lag with respect to Equation 2. These considerations may lead to an improved understanding of how sentiments are influenced through social networks.

References

1. Diddi, P., and Lundy, LK. 2017. Organizational Twitter Use: Content Analysis of Tweets during Breast Cancer Awareness Month. Journal of Health Communication 22 (3): 243–53.
2. Fishbein, M., and Ajzen, I. Belief, Attitude, Intention And Behavior. Addison-Wesley, 1975.
3. Halady, I. and Rao, P. 2010. Does awareness to climate change lead to behavioral change? International Journal of Climate Change Strategies and Management 2 (1): 6-22.
4. Hamed, AA., Ayer, AA., Clark, EM., Irons, EI., Taylor, GT., and Zia, A. 2015. Measuring Climate Change on Twitter Using Google's Algorithm: Perception and Events. International Journal of Web Information Systems 11 (4): 527–44.
5. Hodeghatta, UR., and Sahney, S. 2016. Understanding Twitter as an E-WOM. Journal of Systems and Information Technology 18 (1): 89–115.
6. Reips, U, and Garaizar, P. 2011. Mining Twitter: A Source for Psychological Wisdom of the Crowds. Behavior Research Methods 43 (3): 635–42.
7. Rinker, TW. 2019. sentimentr: Calculate Text Polarity Sentiment version 2.7.1. http://github.com/trinker/sentimentr

Homogeneous Symmetrical Threshold Model with Nonconformity

Bartłomiej Nowak, Katarzyna Sznajd-Weron

Department of Theoretical Physics
Wrocław University of Science and Technology
Wrocław, Poland
bartlomiej.nowak@pwr.edu.pl

Models of opinion dynamics that show discontinuous phase transition are one of the most desirable. One of the main reasons for this may be the existence of so-called social hysteresis in many societies, animal as well as human. Thus we ask a question about the possibility of discontinuous phase transition within threshold models. It was checked for some type of majority vote model [1], but we want to check if discontinuous phase transition is possible, when we introduce non-absolute majority type.

Hence, we analyze two variants of the modified Watts threshold model with a noise [2][3]. Models are analyzed analytically using the Mean-Field Approximation, Pair Approximation method [4] and numerically by Monte Carlo simulations. All models are considered on the complete graph, random regular graph and Watts-Strogatz graph. Agents are affected by two forces, conformity and nonconformity (independence or anticonformity), which can order or disorder the system. Here conformity acts as an ordering force and anticonformity acts in the opposite way, i.e. disorders the system. As an order parameter, we use magnetization which is defined as the mean across all states of agents in the system.

We consider a system of N agents, which are described by the binary variables $S = \pm 1$. Agents are placed in the nodes of an arbitrary graph. At each elementary time step, we pick one agent randomly and decide which of two types of behavior she/he will perform in a given time step: with probability p an agent will nonconform (anticonforms or acts independently) and with probability $1-p$ conform to the major opinion. In both cases, we check if the concentration of agents with opinion $S = \pm 1$ across all neighbors is bigger than a set threshold $r \geqslant 0.5$, i.e. we check which opinion is major in the neighborhood. In case of conformity a voter will take the same state as the major opinion and in the case of anticonformity the state opposite to the major one. In case of independence, an agent acts independently, i.e. with probability $\frac{1}{2}$ flips to the opposite state.

We investigate the model via the mean-field approach, which gives the exact result in the case of a complete graph, as well as via Monte Carlo simulations. General results for the model with independence [3]:

$$p = \frac{c_{st} B_{1-c_{st}} - (1-c_{st}) B_{c_{st}}}{\frac{1}{2} - c_{st} - (1-c_{st}) B_{c_{st}} + c_{st} B_{1-c_{st}}}, \tag{1}$$

whereas for the model with anticonformity:

$$p = \frac{B_{c_{st}} - c_{st}(B_{c_{st}} + B_{1-c_{st}})}{B_{c_{st}} - B_{1-c_{st}}}, \tag{2}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where

$$\begin{aligned} B_c &= P(X_1 \geqslant \lfloor r(N-1) \rfloor), \\ B_{1-c} &= P(X_2 \geqslant \lfloor r(N-1) \rfloor), \end{aligned} \tag{3}$$

where X_1 is a binomially distributed random variable with $N-1$ number of trials and success probability in each trial equal to c, and X_2 is a binomially distributed random variable with $N-1$ number of trials and success probability in each trial $1-c$.

We show that indeed if the threshold $r = 0.5$, which corresponds to the majority-vote model, an order-disorder transition is continuous. Moreover, results obtained for both versions of the model (one with independence and the second one with anticonformity) give the same results, only rescaled by the factor of 2. However, for $r > 0.5$ the jump of the order parameter and the hysteresis is observed for the model with independence, and both versions of the model give qualitatively different results, see Fig. 1

Moreover, similar tendencies were observed on a random regular graph and Watts-Strogatz graph. We observe exactly the same behavior as before for parameter $k = N-1$, what corresponds to the complete graph (k describes degree for all nodes in the network). But additionally we observe some interesting behavior for other values of k, for example, parameter k seems to be responsible for discontinuity of the order parameter in the independence case. In the Watts-Strogatz graph case, parameter β (rewiring probability) seems to change position of tipping point. We check also if we can observe the 1st order phase transition in the anticonformity case on the complete graph. For all networks we derive an analytical solution using Mean-Field and Pair Approximation approach. And as before we have validated them by Monte Carlo simulations.

References

1. J. Encinas, H. Chen, M. M. de Oliveira, and C. E. Fiore, *Majority vote model with ancillary noise in complex networks*, Physica A: Statistical Mechanics and its Applications, vol. 516, pp. 563 - 570, 2019.
2. D. J. Watts, *A simple model of global cascades on random networks*, Proceedings of the National Academy of Sciencesof the United States of America 99 (9) (2002), pp. 5766 - 5771
3. Bartłomiej Nowak and Katarzyna Sznajd-Weron, *Homogeneous Symmetrical Threshold Model with Nonconformity: Independence versus Anticonformity*, Complexity, vol. 2019, Article ID 5150825, 14 pages, 2019. https://doi.org/10.1155/2019/5150825.
4. Jędrzejewski, Arkadiusz, *Pair approximation for the q-voter model with independence on complex networks*, Phys. Rev. E, 95 (1) (2017), p. 012307, 10.1103/PhysRevE.95.012307

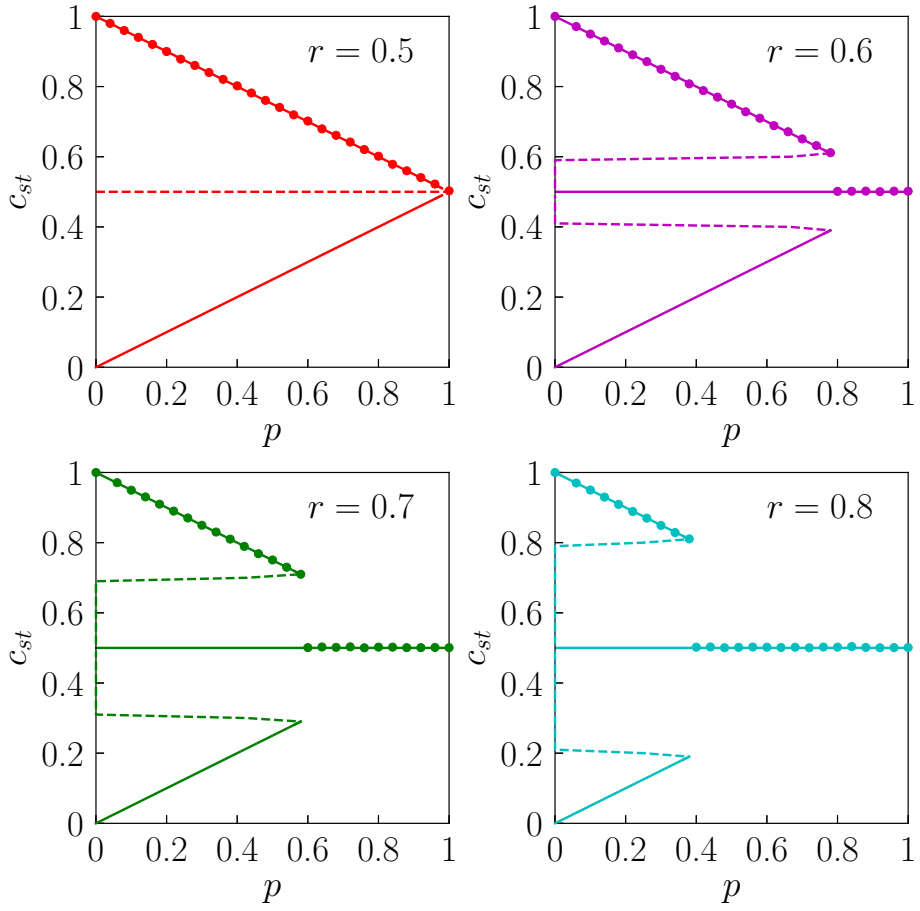


Fig. 1. Phase diagrams for the model with independence for different values of the threshold r. Lines indicate the analytical prediction from MFA and dots represent results of MCS from the initial fully ordered state ($c(0) = 1$) for the system of size $N = 5 \cdot 10^4$ [3].

Semantic Networks and Belief Change

Tamara van der Does[1], Mirta Galesic[1], Nina Fedoroff[2], and Daniel L. Stein[1,3]

[1] Santa Fe Institute, Santa Fe NM 87501, USA,
tamara@santafe.edu
[2] Penn State University, University Park PA 16802, USA
[3] New York University, New York City NY 10003, USA

Many people hold beliefs about scientific issues that are not in line with the scientific consensus. Even though 86% of scientists who are members of AAAS think that parents should be required to vaccinate their healthy children and 88% think genetically modified (GM) food is safe to eat, only 68% of the U.S. public think that all healthy children should be vaccinated and 37% think it is safe to eat GM food [8]. Erroneous beliefs about scientific issues can have important societal consequences, including measles outbreaks [7] and precarious farming economies [6].

Beliefs about scientific issues are often related to various moral considerations in complex semantic networks [1]. For instance, beliefs about vaccines and GM food can be connected to the perceived unfairness of the practices of pharmaceutical and biotech companies, which in turn might be related to environmental concerns. Using these indirect moral arguments could be more effective at changing minds than solely providing facts [9]. In other words, when there is a strong relationship between a scientific belief and a moral consideration, it might be necessary to first change the moral consideration in order to change the scientific belief. However, if this moral consideration is itself tightly associated with other moral considerations in one's semantic network, or structurally embedded [3], it might be difficult to change it and, consequently, to change the related scientific belief [5].

We explore how structural embeddedness of moral considerations related to beliefs about vaccines and GM food affects the likelihood of belief change about the safety of these technologies. We hypothesize that 1) beliefs about a moral consideration related to vaccines and GM food will be less likely to change if that consideration is strongly connected to other moral considerations within one's semantic network, and 2) beliefs about the safety of vaccines and GM food will change most after interventions targeting considerations that are well connected to safety concerns but, at the same time, less connected to other moral considerations.

1 Methods

We collected data within a longitudinal experimental survey with Mechanical Turk participants whose beliefs were not aligned with the scientific consensus about the safety of childhood vaccination ($N = 409$) or GM food ($N = 406$).

In three survey waves, we measured participants' beliefs about safety of vaccines and GM food (safety beliefs), as well as their related moral considerations belonging to

This work is supported by award no. 2018-67023-27677 from the USDA National Institute of Food and Agriculture. The funder had no role in study design or interpretation of the results.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

six different moral domains [4]: whether they benefit children and environment (Care), are part of one's tradition and approved by appropriate agencies (Authority/Respect), are natural and in line with God (Purity/Sanctity), positively affect one's family and country (Loyalty), are fair to different actors, such as big corporations, patients, and farmers (Fairness), and whether one is free to choose these technologies and has access to all important information (Freedom). For each consideration, we also measured how important it is for one's safety beliefs. In the second wave, participants received educational interventions including scientific facts about the safety of these technologies combined with messages targeting one moral consideration, either harm to children, naturalness, or fairness regarding the profit of big companies vs. patients (vaccines) or large vs. small farmers (GM food). The study also included questions and interventions regarding social norms, which will be reported elsewhere.

We constructed a network of moral considerations where edges represented partial correlations of each consideration with others [2]. We calculated correlations between each moral consideration and safety beliefs. We approximated structural embeddedness with two measures: 1) closeness, reflecting the average of each consideration's partial correlations with all other considerations (the network is fully connected); and 2) weighted closeness, where each partial correlation was weighted by the consideration's reported importance for beliefs about safety. The resulting coefficients for closeness ranged from .7 to .9, and for weighted closeness from .38 to .64 for vaccines and GM food, respectively. For each moral consideration, we also computed the ratio of its correlation with the belief about safety over its structural embeddedness. For closeness (weighted closeness), this ratio ranged from .68 to .88 (10 to 21) for vaccines, and from .34 to .70 (7 to 11) for GM food. Finally, we calculated the change in beliefs about safety after educational intervention, as the absolute proportional difference in beliefs reported in Wave 1 and immediately after the intervention in Wave 2, as well as a week later in Wave 3. These changes ranged from 11 and 24 percentage points.

2 Results

Figure 1 shows network of moral considerations related to safety beliefs about vaccination and GM food. In line with our first hypothesis, we find that changes in beliefs about the moral considerations targeted by our educational interventions are strongly negatively correlated with their structural embeddedness for both beliefs about vaccines and GM food, when measured as simple closeness ($r = -.53$ and $r = -.97$, respectively). For weighted closeness, the same relationship holds for GM food ($r = -.86$), but not for vaccines ($r = .69$), where a moral value with high weighted closeness (the profit of big companies) did experience significant changes. In line with our second hypothesis, we find that changes in beliefs about safety are positively correlated with more correlated yet less embedded target considerations. The results hold for structural embeddedness measured both as closeness ($r = .43$ and $r = .35$) and weighted closeness ($r = .48$ and $r = .75$, for vaccines and GM food, respectively).

In sum, we show that structural embeddedness of moral considerations strongly affects the likelihood of changing beliefs about the safety of vaccines and GM food after educational interventions. While the best educational interventions for each scien-

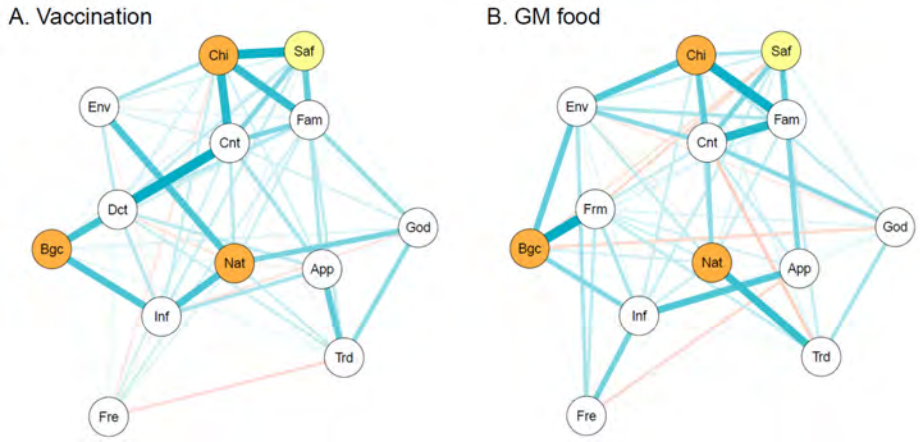


Fig. 1. Relationship of different moral considerations with beliefs about the safety of A. vaccination and B. GM food. Thicker lines indicate stronger partial correlations, nodes in orange represent targeted educational interventions, the yellow node represents the outcome.

tific issue differ in content, for both vaccines and GM the intervention producing most change has one of the highest ratios of its correlation with safety beliefs over structural embeddedness. Our results suggest that to change scientific beliefs, one should first attempt to change underlying moral considerations, focusing on those that are important for the scientific beliefs but not tightly interconnected with other considerations. In further research, we will explore other measures of structural embeddedness, such as betweenness, and study the accuracy of estimated networks of moral considerations.

References

1. Dalege, J., Borsboom, D., Van Harreveld, F., et al.: Toward a formalized account of attitudes: The Causal Attitude Network (CAN) Model. Psychol. Rev. 123(1), 2–22 (2016)
2. Epskamp, S.,Cramer, A.O., Waldorp, L.J., et al.: qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software, 48(4), 1-18 (2012).
3. Granovetter, M.: Economic Action and Social Structure: The Problem of Embeddedness. Am. J. Sociol. 91(3), 481–510 (1985)
4. Haidt, J., Kesebir, S.: Morality. In Fiske, S., Gilbert, D., Lindzey, G. (eds.): Handbook of Social Psychology, pp. 797–832. 5th edn. Wiley, Hobeken, N.J. (2010)
5. Kahan, D. M., JenkinsSmith, H., Braman, D.: Cultural cognition of scientific consensus. Journal of risk research, 14(2), 147-174 (2011).
6. Lucht, J. M.: Public acceptance of plant biotechnology and GM crops. Viruses 7(8), 4254–4281 (2015)
7. Patel, M., Lee, A. D., Redd, S. B., et al.: Increase in Measles Cases - United States, January 1-April 26, 2019. US Department of Health and Human Services/Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report. 68(17), 402–404 (2019)
8. Pew Research Center: Americans, Politics and Science Issues. pp. 1–175 (2015)
9. Steele, C. M., Ostrom, T. M.: Perspective-mediated attitude change: When is indirect persuasion more effective than direct persuasion? Journal of Personality and Social Psychology 29(6), 737–741 (1974)

Part VI

Link Analysis and Ranking

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Axiomatization of the PageRank Centrality

Tomasz Wąs and Oskar Skibski

Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
{t.was,o.skibski}@mimuw.edu.pl

1 Introduction

PageRank, introduced over 20 years ago by Page *et al.* [5], is one of the standard tools for network analysis. While nowadays PageRank is used in various settings, initially it was developed for measuring the importance of webpages on the Internet. Using graph terminology, Internet network is a directed multigraph, were webpages are nodes and links between them—edges. Additionally, the node weights can be understood as a baseline importance of a webpage coming from predefined preferences of the user [5], its relevance to a specific topic [4], or another source.

PageRank belongs to a class of *feedback centrality measures*. These centralities are defined through a recursive formula that links the centrality of a node with the centralities of its direct predecessors. More in detail, PageRank of a node is the sum of two parts: the first one is proportional to the sum of PageRanks of its direct predecessors divided by the number of their outgoing edges; the second one is the weight of this node, i.e., constant b_v. Formally,

$$PR_v(G) = a \cdot \left(\sum_{(u,v)\in E} \frac{PR_u(G)}{\textit{out-degree}(u)} \right) + b_v.$$

In this paper, we analyse PageRank by providing its axiomatic characterisation—a set of simple, intuitive properties that uniquely characterise a centrality measure. This approach allows to highlight similarities and differences between centrality measures and in recent years has gained popularity in the literature, e.g.: to emphasise the usefulness of the *Harmonic Centrality* [3] or to find common patterns in the behaviour of different centrality measures [2, 7]. In our work, we capture the intuition behind PageRank with six simple axioms namely: *Foreseeability*, *Outgoing Homogeneity*, *Edge Swap*, *Sink Merging*, *Twin Sources*, and *Dummy Node*, and we prove that PageRank is the only centrality measure that satisfies all six of them.

This is the first axiomatic characterisation of the PageRank centrality in its original, general form. So far, only simplified version of PageRank (without constant b_v) has been axiomatized: Palacios-Huerta and Volij [6] axiomatized *Invariance Method* which is equivalent to a simplified PageRank in a setting of a scientific journal citation network; In turn, Altman and Tennenholtz [1] focused on the ranking that results from the simplified PageRank.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

Let us present our six axioms and explain the intuition behind them. Five of them define graph operations under which the centrality should be invariant. The sixth axiom, namely *Dummy Node*, determines the exact centrality of a node in the borderline case.

Our first axiom, *Foreseeability*, states that the importance of a webpage depends mostly on its backlinks, not its links. Hence, if we remove the part of a network, from which we cannot reach our webpage by a sequence of links, then the importance of our webpage should still be the same.

> ***Foreseeability:*** *For every graph $G = (V,E)$ and node $v \in V$, removing everything but a subgraph consisted of v, all its predecessors and their outgoing edges does not affect the centrality of v.*

Observe that since the importance of a link on a webpage depends on the total number of its links, the outgoing edges of all predecessors of v have to be preserved.

Outgoing Homogeneity, our next axiom, states that the absolute number of links on a webpage does not impact the importance of any webpage. Since creating a link has practically zero cost, this property is important to prevent ranking manipulations and laid at the foundation of PageRank.

> ***Outgoing Homogeneity:*** *For every graph $G = (V,E)$ and constant $k \in \mathbb{N}$, adding k copies of all outgoing edges of node $v \in V$ does not affect any centrality in the graph.*

For the next axiom, *Edge Swap*, consider the case when there are two equally important webpages with equal number of links. Then, the axiom states that the links from these webpages have equal impact—it does not matter for the importance of any webpage from which of these two webpages it has a backlink. This property is characteristic for feedback centralities.

> ***Edge Swap:*** *For every graph $G = (V,E)$, if nodes $u,v \in V$ have equal centralities and equal number of outgoing edges, then replacing edges $(u,u'),(v,v') \in E$ for edges $(u,v'),(v,u')$ does not affect any centrality in the graph.*

In our next axiom, *Sink Merging*, we focus on a situation, where two webpages are merged into one, preserving their backlinks. If there are no links on both webpages, then after this operation the importance of a merged webpage is a sum of the original importance of both webpages.

> ***Sink Merging:*** *For every graph $G = (V,E)$, merging two sinks $u,v \in V$ does not affect the centralities of the remaining nodes in the graph; moreover, the centrality of the merged node is the sum of the centralities of nodes u and v in graph G.*

Now, imagine that there are two identical webpages without backlinks. In such a case, our next axiom, *Twin Sources*, states that if we remove one of them and transfer its baseline importance to the other webpage, then the importance of webpages in the rest of the network will not be affected.

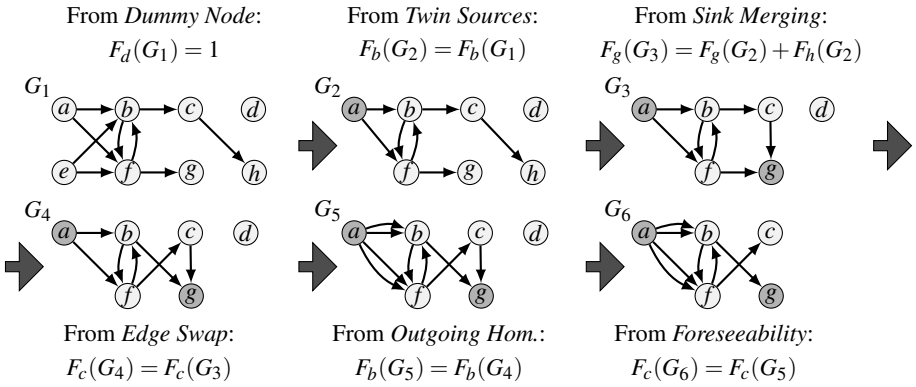


Fig. 1. Six graphs illustrating our axioms. Light grey nodes have weight 1 and dark grey—2.

Twin Sources: *For every graph $G = (V,E)$ and two sources $u, v \in V$ with identical set of edges, removing u and adding its weight to the weight of v does not affect the centralities of the remaining nodes in the graph; moreover, the centrality of v is the sum of the centralities of nodes u and v in graph G.*

In our last axiom, we consider a webpage without any links nor backlinks.

Dummy Node: *For every graph $G = (V,E)$, if node $v \in V$ does not have any edges (outgoing nor incoming), then its centrality is equal to its weight.*

Our main result states that these six axioms uniquely characterise PageRank.

Theorem 1. *PageRank centrality is a unique centrality measure satisfying Foreseeability, Outgoing Homogeneity, Edge Swap, Sink Merging, Twin Sources and Dummy Node.*

The early version of this work was presented at the IJCAI-18 conference [8].

References

1. Altman, A., Tennenholtz, M.: Ranking systems: the PageRank axioms. In: Proceedings of the 6th ACM Conference on Electronic Commerce (ACM-EC). pp. 1–8 (2005)
2. Bloch, F., Jackson, M.O., Tebaldi, P.: Centrality measures in networks (2016), available at SSRN 2749124
3. Boldi, P., Vigna, S.: Axioms for centrality. Internet Mathematics 10(3-4), 222–262 (2014)
4. Haveliwala, T.H.: Topic-sensitive pagerank. In: Proceedings of the 11th international conference on World Wide Web. pp. 517–526. ACM (2002)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford InfoLab (1999)
6. Palacios-Huerta, I., Volij, O.: The measurement of intellectual influence. Econometrica 72(3), 963–977 (2004)
7. Schoch, D., Brandes, U.: Stars, neighborhood inclusion, and network centrality. In: SIAM Workshop on Network Science (2015)
8. Wąs, T., Skibski, O.: Axiomatization of the PageRank centrality. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). pp. 3898–3904 (2018)

Link Prediction in Signed Social Networks: from Status Theory to Motif Families

Jing Xiao, Si-Yuan Liu, and Xiao-Ke Xu

College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China

1 Introduction

Signed networks are a special type of complex networks with both positive and negative edges. The positive edges represent positive relationships such as "friends" and "trust", and are represented by the positive sign "+". The negative edges represent negative relationships such as "enemies" and "distrust", and are represented by the negative sign "–". In a signed network, whether two nodes do link each other depends on not only the number of common neighbors between them, but also the sign and direction of each edge in their neighborhood.

The most commonly used link prediction algorithm in signed networks is based on small subgraphs that satisfy status theory, and these subgraphs can be understood as special cases of motifs [1]. Compared with many global structural features such as small-world and scale-free, motif (i.e., subgraph) is the most basic structural and functional unit in complex networks [2]. Link prediction via a motif can be expressed as: whether two nodes do connect depends on the specific functional units formed by the edge connecting these two nodes and their neighbors [3]. The motif-based prediction algorithm considers the connection patterns (including the sign and direction of edges) between node pairs and their neighbors, so it is applicable to signed networks [4].

The existing motif-based link prediction algorithms for signed networks have following three drawbacks. First, the current methods only focus on motifs that satisfy status theory [5], but do not consider other types of motifs. Actually, the mechanism by which the motifs can be employed to link prediction in signed networks is not explained. At the same time, there is no answer as to explain the mechanism that calculating the number of motifs on the predicted edge can be used for link prediction. Finally, the classical algorithms of link prediction are based on only a single motif and do not think about the relation between different kinds of motifs.

To solve the above mentioned problem, we investigate a novel framework based on edge-dependent motif for link prediction. In this study, we first use motif theory to explore the relationship between the number of each motif and its ability for link prediction. Experiments on five empirical signed networks demonstrate that the prediction ability of a motif depends not on its number in the whole network but on the number of edge-dependent motifs. Then we explain the edge-dependent motif based link prediction by a naive Bayes model. Secondly, we put forward a signed naive Bayes model combining two motifs, which has higher prediction performance than a single motif. Finally, We combine all the types of 3-node motifs to build a machine learning classifier

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

based on motif families. The network structure information used by motif families in link prediction is more comprehensive than status theory and thus gives more accurate prediction.

2 Results

We combine all the predictors for positive edges to construct motif families. Treat the scores of edges calculated by 16 predictors as 16-dimensional features, and then use XGboost for link prediction. The prediction results of the five large-scale signed networks based on all the predictors for positive edges are shown in Tables 1. In each column, the best result and the result based on the motif family (all the motifs) are highlighted in boldface. From these two tables, link prediction using the motif family is more accurate than using a single motif, and this conclusion can be drawn from all the five experimental networks. The motif families not only consider the motifs that satisfy status theory, but also utilize the motifs that do not satisfy status theory, so they have higher prediction performance.

Table 1. The results of link prediction by combining multiple positive predictors. Here P represents the result of Precision.

Motif	Bitcoinalpha		Bitcoinotc		Wiki-RfA		Slashdot		Epinions	
	AUC	P	AUC	P	AUC	P	AUC	P	AUC	P
S_1	0.782	**0.996**	0.775	0.997	0.814	0.988	0.634	0.999	0.838	**1.000**
S_2	0.780	0.993	0.774	0.996	0.775	0.988	0.634	0.998	0.821	**1.000**
S_3	**0.786**	0.996	**0.778**	**0.997**	**0.913**	**0.996**	**0.655**	**1.000**	**0.841**	**1.000**
S_4	0.534	0.902	0.533	0.840	0.608	0.874	0.515	0.661	0.543	0.708
S_5	0.509	0.605	0.511	0.611	0.515	0.548	0.506	0.562	0.524	0.614
S_6	0.509	0.613	0.517	0.679	0.563	0.714	0.533	0.843	0.599	0.974
S_7	0.533	0.878	0.529	0.823	0.650	0.972	0.527	0.788	0.582	0.892
S_8	0.539	0.951	0.537	0.879	0.641	0.960	0.523	0.739	0.560	0.786
S_9	0.548	0.966	0.542	0.937	0.555	0.692	0.510	0.608	0.548	0.731
S_{10}	0.539	0.929	0.533	0.839	0.547	0.662	0.511	0.614	0.564	0.807
S_{11}	0.537	0.916	0.533	0.842	0.624	0.933	0.521	0.718	0.555	0.763
S_{12}	0.550	0.981	0.543	0.945	0.706	0.976	0.522	0.733	0.540	0.692
S_{13}	0.777	0.995	0.766	0.998	0.638	0.935	0.572	0.985	0.736	0.999
S_{14}	0.508	0.585	0.510	0.613	0.554	0.689	0.515	0.663	0.519	0.593
S_{15}	0.533	0.868	0.529	0.810	0.613	0.893	0.518	0.697	0.541	0.697
S_{16}	0.539	0.933	0.536	0.896	0.637	0.962	0.517	0.682	0.530	0.644
All	**0.823**	**0.996**	**0.825**	**0.998**	**0.959**	**0.998**	**0.746**	**1.000**	**0.899**	**1.000**

Then, we compare our proposed method (i.e., Motif Family) with two state-of-the-art methods in signed networks: FriendTNS [6, 7] and Status Theory [5]. The results of the predictors for positive edges are shown in Fig. 1, as the size of the training set increases, the predictive performance of all methods is improved. Furthermore, motif-based methods (i.e., Motif Family and Status Theory) can obtain higher prediction results than FriendTNS, and the method of Motif Family obtains the best predictive performance because it considers more types of motifs (i.e., motifs that do not satisfy status theory) than Status Theory.

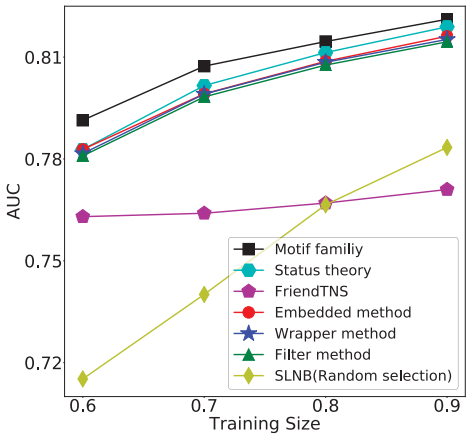


Fig. 1. The comparison of our proposed method with the existing methods.

References

1. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science **298**(5594) (2002) 824–827
2. Xu, X., Zhang, J., Michael, S.: Superfamily phenomena and motifs of networks induced from time series. Proceedings of the National Academy of Sciences of the United States of America **105**(50) (2008) 19601–19605
3. Aghabozorgi, F., Khayyambashi, M.R.: A new similarity measure for link prediction based on local structures in social networks. Physica A: Statistical Mechanics and its Applications **501** (2018) 12–23
4. Zhang, Q., Lü, L., Wang, W., Zhou, T.: Potential theory for directed networks. Plos One **8**(2) (2013) e55437
5. Li, X.: Towards practical link prediction approaches in signed social networks. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. (2018)
6. Symeonidis, P., Tiakas, E., Manolopoulos, Y.: Transitive node similarity for link prediction in social networks with positive and negative links. In: Acm Conference on Recommender Systems. (2010)
7. Symeonidis, P., Tiakas, E.: Transitive node similarity: Predicting and recommending links in signed social networks. World Wide Web **17**(4) (2014) 743–776

A novel measure of edge and vertex centrality for assessing robustness in complex networks

G.P.Clemente[1] and A.Cornaro[2]

[1] Dipartimento di Matematica per le Scienze Economiche, Finanziarie e Attuariali
Universitá Cattolica del Sacro Cuore, Milano
gianpaolo.clemente@unicatt.it
[2] Dipartimento di Matematica per le Scienze Economiche, Finanziarie e Attuariali
Universitá Cattolica del Sacro Cuore, Milano
alessandra.cornaro@unicatt.it

1 Introduction

Network structures are pervasive around us and numerous critical facilities are connected together by various networks. The function and behaviour of networked systems can be largely influenced by their structural features. In such a framework, network topology plays an important role. A fundamental issue concerning complex networked systems is indeed the robustness of the overall system to the failure of its constituent parts (see [1]). In the recent literature a graph measure called *Effective graph resistance*, also known as *Kirchhoff index*, has gained increasing attention in network robustness theory. This topological indicator is defined as the accumulated effective resistance between all pairs of vertices (see [5]) and it has been reformulated as a function of the real eigenvalues μ_i of the Laplacian matrix associated to a graph (see [3])

$$K(G) = n \sum_{i=1}^{n-1} \frac{1}{\mu_i}. \tag{1}$$

In order to compare the value of the Kirchhoff index for networks with a different number of vertices n, we can consider the *normalized Kirchhoff index* defined as (see [6]):

$$K^N(G) = \frac{K(G)}{\binom{n}{2}}, \tag{2}$$

where the denominator considers the maximum number of edges[3].
In [2] and [7], the authors showed that the Kirchhoff index is suitable for assessing the ability of a network to continue performing well when it is subject to failure and/or attack. In fact, the pairwise effective resistance measures the vulnerability of a connection between a pair of vertices that considers both the number of paths between the vertices and their length. A small value of the effective graph resistance therefore indicates a robust network. A very interesting feature of the Kirchhoff index is that it

[3] As an alternative normalization, for sparse graphs, it is possible to consider the number of vertices.

shows strict monotonicity when edges are added or removed, in particular it strictly decreases/increases when edges are added/removed. However, it remains a challenge to identify a specific indicator that displays all the desirable properties usually requested for a robustness quantifier and that can be functional to evaluate and compare real-world networks, especially when topological changes in the network structure have been occurred. Moving along this line, we aimed at presenting a novel robustness measure, which we refer to as *Effective Resistance Centrality*, based on the Kirchoff index. In particular the Effective Resistance Centrality of a vertex (or an edge) is defined as the relative drop of the Kirchhoff index caused by the deactivation of this vertex (edge) from the network. In this way, we provide a local robustness measure able to catch which is the effect of either a specific vertex or a specific edge on the network robustness. Since the degree to which a networked system continues to function typically depends on the integrity of the underlying network, the question of system robustness is usually addressed by analysing how the network structure changes as vertices (or edges) are removed. Several works deal with this topic by evaluating the effect on the network structure of vertices removed either randomly (see, e.g., [1]) or on the basis of targeting criteria related to specific centrality measures (see, e.g., [4]). To this end, we provide a new local measure of importance that can be used as a new criterion for node (or edge) selection when targeted attack strategies are implemented. We further investigate the validness of our proposal on a wide variety of well-known model networks and on the United States domestic airport network. In particular, we investigate the role and significance that airports play in maintaining the structure of the entire domestic airport network.

2 Effective Resistance Centrality

We now provide a definition and a structural description of the edge-based Effective Resistance Centrality and vertex-based Effective Resistance Centrality, respectively.

2.1 Edge-based Effective Resistance Centrality

Let $G = (V,E)$ be a k-edge-connected graph (with $k > 1$) of n vertices and m edges and $G_{e_{i,j}}$ the graph obtained by removing the edge $e_{i,j}$, connecting vertices i and j, from G.

Lemma 1. *If $G_{e_{i,j}}$ is an arbitrary subgraph of a graph G, then $K(G_{e_{i,j}}) \geq K(G)$*

It is noteworthy to say that, if G is a 1-edge-connected graph, then the resulting subgraph $G_{e_{i,j}}$ can be disconnected. In this case, when $e_{i,j}$ is a bridge, we have that $K(G_{e_{i,j}}) = \infty$.

We now introduce a new measure able to capture the relevance of an edge in the network and we refer to it as Effective Resistance Centrality. It is mainly based on the idea that the importance (or centrality) of an edge is related to the ability of the network to continue performing well after the deletion of this edge.

Definition 1. *The Effective Resistance Centrality $R_K(e_{i,j},G)$ of the edge $e_{i,j}$ is defined as*

$$R_K(e_{i,j},G) = \frac{(\Delta K)_{e_{i,j}}}{K(G)} = \frac{K(G_{e_{i,j}}) - K(G)}{K(G)}. \tag{3}$$

By Lemma 1, $(\Delta K)_{e_{i,j}} = K(G_{e_{i,j}}) - K(G)$ must be non-negative, therefore, $R_K(e_{i,j}, G)$ displays monotonicity with respect to edge removal.

2.2 Vertex-based Effective Resistance Centrality

Let $G = (V, E)$ be a connected graph of n vertices and m edges and G_{v_i} the graph obtained by removing the vertex v_i and all its related connections from G.

Definition 2. *The Effective Resistance Centrality $R_K(v_i, G)$ of the vertex v_i is defined as*

$$R_K(v_i, G) = \frac{(\Delta K^N)_{v_i}}{K^N(G)} = \frac{K^N(G_{v_i}) - K^N(G)}{K^N(G)}. \tag{4}$$

$R_K(v_i, G)$ is defined by considering at the numerator the drop of the normalized Kirchhoff index. This choice is justified by the fact that we want to provide a consistent comparison between graphs G and G_{v_i} that have different orders. Notice that in Definition 1 an eventual use of the normalized Kirchhoff index would lead to the same results as in (3).
The quantity $(\Delta K^N)_{v_i}$ is not always positive, depending on the relevance of the specific vertex v_i in the network. On one hand, this measure can be useful in order to detect strategic nodes, whose failure can affect the resilience of the network. On the other hand, the measure also allows to identify eventual nodes to be removed in order to improve the robustness of the network.

3 Main numerical results

In the numerical analysis we exploit how node and edges removals affect network classes with different underlying mechanism. In the Erdős-Rényi (ER) graphs, we observe the presence of specific vertices, whose removal can significant improve the robustness of the network. In general, a lower probability of attachment, and therefore, a lower density, leads to increase the index, providing a more vulnerable graph. As well-known, as a random network gets denser, the critical threshold, at which a complex network will lose its giant component, increases, meaning a higher fraction of the nodes must be removed to disconnect the giant component. Concerning Watts and Strogatz (WS) graphs with higher densities, we derive results in line with the ER random graphs. Instead, when low densities are considered, the WS graph appears more vulnerable than ER to random nodes or edges removal. In the Barabási-Albert model, our findings validate the results of recent studies. Although, in general, scale-free networks are extremely resilient to random failures, they are also extremely vulnerable to targeted attacks.
Finally, we explored the behaviour of our proposal by using the peculiar business network of the U.S. airport, where vertices are the airports and edges are related to the presence of at least a domestic flight scheduled among them in 2017. Results show the effectiveness of the measures we propose in catching the peculiar characteristics of different nodes in the airport network. In particular, focusing on large and medium hubs, we are able to emphasize their strategic role in the airport system. We also provided

a consistent comparison with several well-known topological measures that assess the node importance (namely, degree, clustering coefficient, betweenness, closeness, eigenvector centrality and number of passengers). On average, we notice a significant positive dependence between different indicators. For instance, top strategic airports are also selected by the betweenness. It is easy to understand that these airports are vital to the network and pose serious risks to the structure if disrupted. On the other hand, we observe that Effective Resistance Centrality and betweenness rank in a different way medium airports whose removal does not lead to a disconnected graph.

References

1. R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.
2. W. Ellens, F.M. Spieksma, P. Van Mieghem, A. Jamakovic, and R.E. Kooij. Effective graph resistance. *Linear algebra and its applications*, pages 2491–2506, 2011.
3. L. Feng, I. Gutman, and L. Yu. Degree resistance distance of unicyclic graphs. *Trans. Comb.*, 1:27–40, 2010.
4. S. Iyer, T. Killingback, B. Sundaram, and Z. Wang. Attack robustness and centrality of complex networks. *PloS one*, 8(4):e59613, 2013.
5. D. J. Klein and M. Randić. Resistance Distance. *J. Math. Chem.*, 12:81, 1993.
6. H. Wang, H. Hua, and D. Wang. Cacti with minimum, second-minimum, and third-minimum Kirchhoff indices. *Mathematical Communications*, 15:347–358, 2010.
7. X. Wang, E. Pournaras, R.E. Kooij, and P. Van Mieghem. Improving robustness of complex networks via the effective graph resistance. *The European Physical Journal B*, 87(9):221, 2014.

Part VII

Machine Learning and Networks

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

A Framework for Comparing Graph Embeddings

Bogumił Kamiński[1], Paweł Prałat[2], and François Théberge[3]

[1] Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland,
bogumil.kaminski@sgh.waw.pl
[2] Department of Mathematics, Ryerson University, Toronto, ON, Canada,
pralat@ryerson.ca
[3] Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada,
theberge@ieee.org

1 Introduction

A graph embedding is a mapping of the vertices of a graph into k-dimensional vectors. Good embeddings should capture the graph topology and vertex-to-vertex relationship. Several graph embedding algorithms are available and for each algorithm, parameters need to be set such as the dimension of the embedding space. As a result, selecting the best embedding is a challenging task and very often requires domain experts.

We propose an unsupervised framework for the computation of divergence scores to compare the quality of different embeddings for a given graph, where quality is defined as preserving the community structure, as in [2]. The framework relies on two main ingredients: (i) a good, stable graph clustering algorithm; we use the ECG algorithms detailed in [5], and (ii) a generalization of the Chung-Lu model for graphs which incorporates the geometry provided by the graph embedding.

In order to validate our framework, we ran a large number of experiments with synthetic networks as well as real-world networks, using various embedding algorithms.

2 Geometric Chung-Lu Model

In the Chung-Lu model [1], given some degree distribution $\mathbf{w} = (w_1, \ldots, w_n)$ over n vertices $v_1, \ldots, v_n$, edge probabilities of a generated graph are defined such that the expected degrees for the vertices agree with this distribution.

In our proposed *Geometric Chung-Lu* model (GCL), we also consider an embedding of the vertices of G in some k-dimensional space $\mathscr{E} : V \to \mathbb{R}^k$. In particular, for each pair of vertices, v_i, v_j, we know their distance: $\text{dist}(\mathscr{E}(v_i), \mathscr{E}(v_j))$. We consider $0 \leq d_{i,j} \leq 1$, a normalized version of those distances.

In the GCL model, the probability that v_i and v_j are adjacent is proportional to $s(d_{i,j})$, a decreasing function s. For some choice of $\alpha \in [0, \infty)$, we define $s(d_{i,j}) := (1 - d_{i,j})^{\alpha}$ for all $d_{i,j}$'s. This choice gives us a good variety of functions to choose from. Choosing a large value for α makes it less probable to have long edges in embedded space. With a small value for α, the distance in embedded space has less importance, and it is completely ignored when $\alpha = 0$.

The GCL model is the random graph $\mathscr{G}(\mathbf{w}, \mathscr{E}, \alpha)$ on the vertex set $V = \{v_1, \ldots, v_n\}$ in which each pair of vertices v_i, v_j, independently of other pairs, forms an edge with

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

probability $p_{i,j}$, where $p_{i,j} = x_i x_j s(d_{i,j})$ for some learned weights $x_i \in \mathbb{R}_+$. The weights are such that the expected degree of v_i is $w_i = \deg_G(v_i)$ for all $1 \leq i \leq n$.

We show in [3] that there exists a unique selection of weights x_i, provided that the maximum degree of G is less than the sum of degrees of all other vertices. Moreover, we show how to efficiently compute those weights numerically to any desired precision, which can be made even faster via sampling.

3 The Framework

Given a graph $G = (V,E)$, its degree distribution $\mathbf{w}$ on V, and an embedding $\mathscr{E} : V \to \mathbb{R}^k$ of its vertices in k-dimensional space, we perform the five steps detailed below to obtain $\Delta_{\mathscr{E}}(G)$, a *divergence score* for the embedding. We can apply this algorithm to compare several embeddings $\mathscr{E}_1, \ldots, \mathscr{E}_m$, and select the best one via $\text{argmin}_i \Delta_{\mathscr{E}_i}(G)$

Step 1: Run some stable *graph* clustering algorithm on G to obtain a partition $\mathbf{C}$ of the vertex set V into ℓ communities $C_1, \ldots, C_\ell$. We use the ensemble clustering algorithm for graphs (ECG) [5], but any other good algorithm can be used.

Step 2: For each $1 \leq i \leq \ell$, let c_i be the proportion of edges of G with both endpoints in C_i. Similarly, for each $1 \leq i < j \leq \ell$, let $c_{i,j}$ be the proportion of edges of G with one endpoint in C_i and the other one in C_j. Let:

$$\bar{\mathbf{c}} = (c_{1,2}, \ldots, c_{1,\ell}, c_{2,3}, \ldots, c_{2,\ell}, \ldots, c_{\ell-1,\ell}) \quad \hat{\mathbf{c}} = (c_1, \ldots, c_\ell) \tag{1}$$

be two vectors that sum to one. These *graph-based vectors* characterize the partition $\mathbf{C}$ from the perspective of G. The embedding $\mathscr{E}$ does *not* affect these vectors.

Step 3: For a given parameter $\alpha \in \mathbb{R}_+$ and the same vertex partition $\mathbf{C}$, consider $\mathscr{G}(\mathbf{w}, \mathscr{E}, \alpha)$, the GCL model. For each $1 \leq i < j \leq \ell$, we compute $b_{i,j}$, the expected proportion of edges of $\mathscr{G}(\mathbf{w}, \mathscr{E}, \alpha)$ with one endpoint in C_i and the other one in C_j. Similarly, for each $1 \leq i \leq \ell$, let b_i be the expected proportion of edges within C_i. We obtain two vectors:

$$\bar{\mathbf{b}}_{\mathscr{E}}(\alpha) = (b_{1,2}, \ldots, b_{1,\ell}, b_{2,3}, \ldots, b_{2,\ell}, \ldots, b_{\ell-1,\ell}) \quad \hat{\mathbf{b}}_{\mathscr{E}}(\alpha) = (b_1, \ldots, b_\ell) \tag{2}$$

that sum to one. These *GCL-based vectors* characterizes partition $\mathbf{C}$ from the perspective of the embedding $\mathscr{E}$.

Step 4: We use the Jensen-Shannon divergence [4] (JSD) to measure the dissimilarity between the vectors obtained in (1) and (2). In our implementation, we used a simple average, that is,

$$\Delta_\alpha = \frac{1}{2} \cdot \left(JSD(\bar{\mathbf{c}}, \bar{\mathbf{b}}(\alpha)) + JSD(\hat{\mathbf{c}}, \hat{\mathbf{b}}(\alpha)) \right). \tag{3}$$

Step 5: Run steps 3 and 4 for several choices of α; we tried several values on a grid in the range $0 \leq \alpha \leq 10$ in our experiments. Let $\hat{\alpha} = \text{argmin}_\alpha \Delta_\alpha$. We define the *divergence score* for embedding $\mathscr{E}$ on G as: $\Delta_{\mathscr{E}}(G) = \Delta_{\hat{\alpha}}$.

To compare several embeddings of the same graph G, we repeat steps 3-5 above and compare the divergence scores (the lower score, the better). Steps 1-2 are done only once, so we use the same partition into ℓ communities for each embedding.

4 Illustration

We illustrate our framework on the well-known Zachary's Karate Club graph [6]. We generated over 600 embeddings in dimension 2 to 128, using several different algorithms. In Figure 1, we display the best and worst embeddings according to our framework. Projection in 2 dimensions is obtained with UMAP[1]. The different colors and shapes for the vertices correspond to the two known communities in this graph. We clearly see that the best embedding does a much better job at keeping the vertices within each community close. Results over several other real and artificial graphs as well as using different graph clustering algorithms can be found in [3], all with conclusions similar to Figure 1.

Fig. 1. The Karate Club Graph. We show the best (left) and worst embeddings according to our framework given over 600 different choices. Vertex color and shape correspond to the two known communities. We also display the edges from the graph. We clearly see that the best embedding does a much better job at keeping the vertices within each community close.

References

1. Chung F., Lu L. Complex Graphs and Networks. American Mathematical Society (2006).
2. Keikha M.M., Rahgozar M., Asadpour M., Community Aware Random Walk for Network Embedding, pre-print, arXiv:1710.05199 (2018).
3. Kaminski B., Pawel P., Théberge F., An Unsupervised Framework for Comparing Graph Embeddings, pre-print, arXiv:1906.04562 (2019).
4. Lin J. Divergence measures based on the Shannon entropy. In: IEEE Transactions on Information Theory, 37(1), pp.145-151 (1991).
5. Poulin V., Théberge F., Ensemble Clustering for Graphs: Comparison and Applications, Applied Network Science vol. 4, no. 51 (2019).
6. Zachary W., An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33, 452-473 (1977).

[1] `https://github.com/lmcinnes/umap`

Network Embedding For Link Prediction: The Pitfall and Improvement

Xiao-Ke Xu, Ren-Meng Chao, and Jing Xiao

College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China

1 Introduction

The existing link prediction algorithms can be divided into two categories: structural similarity algorithms in network domain [1] and network embedding algorithms in the field of machine learning [2]. In the algorithms of structural similarity, local similarity indices are the most commonly used because of their low computational complexity, such as Common Neighbor (CN) and Local Path (LP) indices [1].

Recently, the technique of network embedding has been widely applied in link prediction[3]. It aims to map network data into a low dimensional space in which the network neighborhood information is maximumly preserved [2]. By representing nodes in a network as vectors, a wide variety of machine learning algorithms can be used to provide a standard, general and effective solution for link prediction [4]. Besides, network embedding has benefited many applications like network visualization, node classification and node clustering. However, in the current researches, there is a lack of systematic comparison of the two algorithms (structural similarity versus network embedding), and few of them study the shortcomings of network embedding algorithms.

2 Results

Table 1 displays the AUC results of these algorithms. In each column, the best result is highlighted in bold. A short-path network refers to the network where the shortest path length between most pairs of nodes is short. On the contrary, A long-path network refers to the network where the shortest path length between most pairs of nodes is long. It can be seen that for short-path networks (i.e., Ht09 , Email, and WC networks), structure similarity algorithms (e.g., CN, LP, and CCN) can achieve excellent predictive performance. Furthermore, structure similarity algorithms significantly outperform network embedding algorithms in short-path networks. Conversely, for long-path networks, the best predictive performance is obtained from network embedding algorithms (e.g., LargeVis and Node2Vec). In other words, network embedding algorithms have a great deficiency when performing link prediction in short-path networks.

To explain the phenomenon that network embedding algorithms have a pitfall when performing link prediction in short-path networks, six real networks are embedded into vector spaces and the Euclidean distances of node pairs are calculated. Figure 1 shows the distance distributions of existent and nonexistent links in different networks, and

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Table 1. Comparison of the performance measured by AUC, results averaging over 10 systematic experiments. $\langle d \rangle$ denotes the average shortest distance.

AUC	Ht09	Email	WC	Power	BP10	MN
CN	**0.776**	**0.932**	0.171	0.615	0.600	0.528
LP	0.757	0.920	**0.953**	0.689	0.695	0.553
CCN	0.760	0.910	0.912	0.872	0.915	0.736
Node2Vec	0.531	0.546	0.667	0.863	0.938	0.801
LargeVis	0.506	0.478	0.556	**0.933**	**0.966**	**0.842**
LINE	0.618	0.730	0.919	0.613	0.466	0.486
GraphWave	0.484	0.627	0.732	0.507	0.541	0.587
$\langle d \rangle$	1.65	1.96	2.22	18.98	20.85	35.3

three short-path networks with lower average shortest distance and three long-path networks with larger average shortest distance are shown in Fig. 1(a)-(c) and Fig. 1(d)-(f), respectively. It is found that in short-path networks, the distributions of existent and nonexistent links overlap to a large extent, which can sharply reduce the algorithmic performance. Conversely, in long-path networks, the distances of existent links are mainly between 2 and 6, while the distances of nonexistent links are mainly between 6 and 8. These two types of links are highly distinguishable, thus better predictive performance can be obtained in these networks.

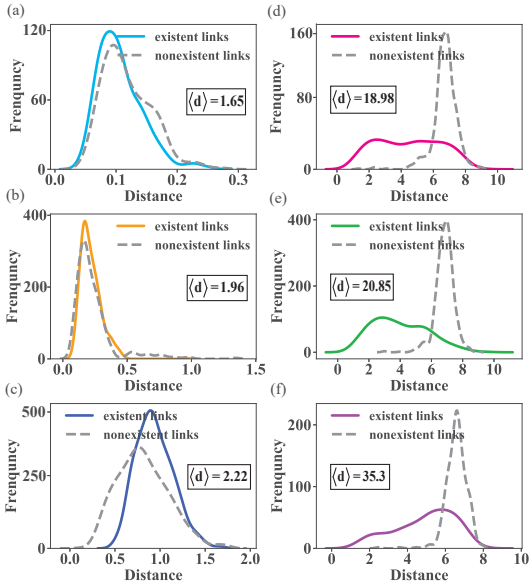


Fig. 1. The Euclidean distance distributions of node pairs in the vector space after network embedding. (a) Ht09, (b) Email, (c) WorldCites, (d) Power, (e) Bcspwr10 and (f) Minnesota.

Based on the above facts, we propose a novel link prediction method to improve the performance of network embedding algorithms, namely, Network Embedding Supplement the information in the Network Domain (NESND), which supplements local structure information with network embedding algorithm and is defined as

$$S_{NESND} = S_{NE} + \lambda S_{ND}, \tag{1}$$

where S_{NE} denotes the network embedding information represented by the Euclidean distance of node pairs, and S_{ND} denotes local structure information from network domain. S_{NESND} denotes the combined information, and λ is a parameter that adjusts how much local structure information is added.

When $\lambda = 1$, Table 2 list all the AUC values. In each column, the best results are highlighted in bold. From the table, it can be seen that for short-path networks, the performance improvements brought by the introduction of CN and LP are greater than CCN index. This is because in short-path networks, compared with community structure information, the number of common neighbor and 3-order paths can more accurately characterize node similarity. By contrast, for long-path networks, especially in Power and MN networks, the enhanced performance brought by CCN index is more significance than CN and LP, because it can predict links accurately in both short-path and long-path networks. The proposed method has 0.2%∼8.3% improvement in long-path networks, while 36.7%∼94.4% improvement can be obtained in short-path networks.

Table 2. Comparison of the performance qualified by the AUC results.

AUC	Ht09	Email	WC	Power	BP10	MN
Node2vec	0.531	0.546	0.667	0.842	0.934	0.801
LargeVis	0.506	0.478	0.556	0.933	0.966	0.842
Node2vec+CN	**0.774**	0.927	0.215	0.850	0.937	0.801
LargeVis+CN	0.767	0.926	0.177	0.935	0.970	0.845
Node2vec+LP	0.772	**0.929**	**0.953**	0.856	0.939	0.801
LargeVis+LP	0.772	**0.929**	**0.953**	0.939	**0.972**	0.848
Node2vec+CCN	0.772	0.920	0.912	0.912	0.958	0.834
LargeVis+CCN	0.772	0.920	0.912	**0.942**	0.968	**0.859**
$\langle d \rangle$	1.65	1.96	2.22	18.98	20.85	35.3

References

1. Ren, Z.M., Zeng, A., Zhang, Y.C.: Structure-oriented prediction in complex networks. Phys. Rep. **750** (2018) 1–51
2. Cai, H.Y., Zheng, V.W., Chang, K.C.C.: A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Trans. Knowl. Data Eng. **30**(9) (2018) 1616–1637
3. Brochier, R., Guille, A., Velcin, J.: Link prediction with mutual attention for text-attributed networks. In: Companion Proceedings of The 2019 World Wide Web Conference. WWW '19, New York, NY, ACM (2019) 283–284
4. Cui, P., Wang, X., Pei, J., Zhu, W.W.: A survey on network embedding. IEEE Trans. Knowl. Data Eng. **31**(5) (2018) 833–852

Optimising the angular coordinates in the hyperbolic embedding of complex networks

Bianka Kovács[1] and Gergely Palla[2]

[1] Dept. of Biological Physics, Eötvös University, Budapest, Hungary,
kovacsbianka@caesar.elte.hu
[2] MTA-ELTE Statistical and Biological Physics Research Group,
Eötvös University, Budapest, Hungary, pallag@hal.elte.hu

1 Introduction

A remarkable network model offering a scale-free degree distribution, high clustering and the small world property at the same time is given by the popularity-similarity-optimization (PSO) model [1]. In this approach the nodes are placed one by one on the Poincaré disk representation of the 2D hyperbolic plane with a logarithmically increasing radial coordinate and a random angular coordinate, and links are introduced with probabilities following the hyperbolic distance between the nodes. The success of the PSO model provides a strong motivation for the development of hyperbolic embedding algorithms, that tackle the inverse problem of finding the optimal hyperbolic coordinates of the nodes based on the network structure. One of the very promising approaches to hyperbolic embedding is given by the noncentered minimum curvilinear embedding (ncMCE) method [2, 3], offering a high quality embedding at a low running time. In the present work we propose a further optimisation of the angular coordinates in the framework of the ncMCE approach that seems to reduce further the logarithmic loss of the embedding compared to the original version, thereby adding an extra improvement to the quality of the inferred hyperbolic coordinates.

2 Methods

In vague terms, the degree of nodes in the PSO model is determined by their radial coordinate (lower distance from the origin corresponds to larger degree), and the angular proximity of the nodes can be interpreted as a sort of similarity, where more similar nodes have a higher probability to be connected. Therefore, in most embedding algorithms the radial coordinates are determined based on the degree of the nodes, whereas the angular coordinates are obtained from some optimisation. In the case of the ncMCE, first a minimum curvilinear distance matrix is prepared, which is then subjected to singular value decomposition, and the angular coordinates are obtained from the vectors corresponding to the first two largest singular values [2, 3].

To measure the quality of the obtained coordinates we can use the logarithmic loss, corresponding to the log-likelihood of the observed adjacency matrix $\boldsymbol{A}$ given the

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

hyperbolic coordinates $\boldsymbol{X}$, written as

$$LL(\boldsymbol{X}) \equiv -\ln\mathcal{L}(\boldsymbol{A}|\boldsymbol{X}) = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} A_{ij}\cdot\ln(p(x_{ij})) - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(1-A_{ij})\cdot\ln(1-p(x_{ij})), \tag{1}$$

where N denotes the number of nodes and the linking probability $p(x_{ij})$ depending on the hyperbolic distance x_{ij} between nodes i and j is specified by the PSO model as $p(x) \approx (1+e^{\frac{\zeta}{2T}(x-R_N)})^{-1}$, where ζ and T are model parameters and R_N is a system size dependent radius.

Our suggestion for improving the angular coordinates obtained from the ncMCE is a direct optimisation of the logarithmic loss, by iterating over the nodes, and in each iteration optionally modifying the angular coordinate of the current node by calculating the logarithmic loss for a couple of new angular positions. If a lower LL can be achieved compared to the original one, the angular coordinate of the node is changed. Since the original angular coordinates given by the ncMCE algorithm are already quite good, we can restrict the search for new angular coordinates within the second angular neighbours of the nodes. The advantage of this choice is that it also allows swaps in the angular order. The number of tried new angular positions per node, q and the total number of correction rounds, n are parameters of our method, which of course, should be kept as low as possible for efficiency. Due to the additive form of (1), at fixed q and n, the time complexity of our algorithm is proportional to N^2.

3 Results

We tested the proposed angular optimisation on both synthetic networks generated with the PSO model and real networks. According to our results, by working with $q = 6$ new angular positions per node, $LL(\boldsymbol{X})$ can be decreased roughly by 15-20% on average during the first 5 to 10 rounds. In Fig. 1a we show the average of LL as a function of the system size for synthetic networks, whereas in Fig. 1b we plot the relative improvement in LL as a function of the number of correction rounds n under the same settings. The curves show that the angular optimisation can provide a significant decrease in LL for both small and larger networks, and the relative improvement seems to converge to a steady value already at $n = 6-8$ rounds.

In Fig. 2 we show the results for a food web among $N = 142$ Cambrian species in the Burgess Shale [4]. The relative improvement in the logarithmic loss (displayed in Fig. 2a) seems to converge again only under 10 rounds to a value close to 15%. In Fig. 2b we show the layout of the network on the Poincaré disk. According to the figure, the originally homogeneous angular arrangement provided by the ncMCE algorithm has become inhomogeneous, allowing a more clear separation between groups of nodes which roughly match the trophic roles of the species.

In conclusion, the proposed optimisation of the angular coordinates seems to provide a substantial reduction in the logarithmic loss of the embedding, at the cost of a relatively low number of extra rounds of iterations over the nodes. In addition, the modification of the coordinates for the studied real network seemed to be quite useful. Based on these,

our extension to the ncMCE can be beneficial in further practical applications where high quality hyperbolic embedding of networks is important.

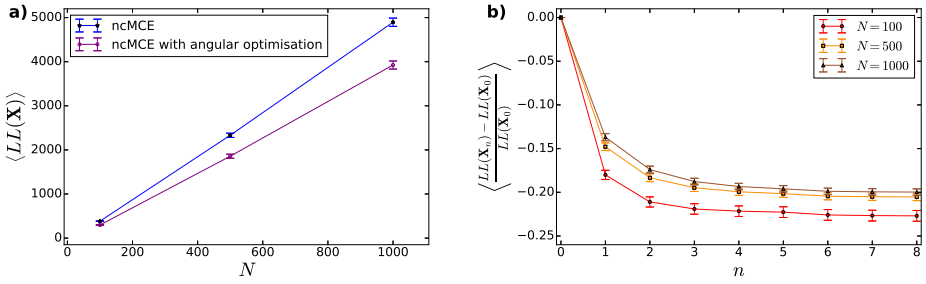


Fig. 1. Logarithmic loss for synthetic networks. a) The average of $LL(\boldsymbol{X})$ for the original ncMCE algorithm (blue) and the ncMCE with angular optimisation (purple) as a function of the number of nodes N, for 100 networks generated by the PSO model with input parameters $\zeta^2 = 1$, $m = 2$, $\beta = 2/3$ and $T = 0.3$. b) The relative improvement in LL as a function of n.

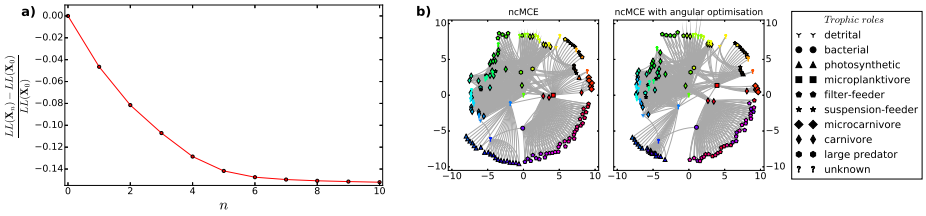


Fig. 2. Results for the food web among Cambrian species in the Burgess Shale. a) The relative improvement in LL as a function of the number of correction rounds n. b) The layout of the network on the Poincaré disk resulted from the original ncMCE algorithm (left) and the ncMCE with angular optimisation (right). The spatial separation of the nodes according to their trophic roles is clearly visible in the hyperbolic plane.

References

1. Papadopoulos F, Kitsak M, Serrano MÁ, Boguñá M, Krioukov D. Popularity versus similarity in growing networks. Nature. 2012;489:537–540.
2. Cannistraci CV, Alanis-Lobato G, Ravasi T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. Bioinformatics. 2013 06;29(13):i199–i209. Available from: https://doi.org/10.1093/bioinformatics/btt208.
3. Muscoloni A, Thomas JM, Ciucci S, Bianconi G, Cannistraci CV. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. Nature Communications. 2017;8:1615.
4. Dunne JA, Williams RJ, Martinez ND, Wood RA, Erwin DH. Compilation and Network Analyses of Cambrian Food Webs. PLOS Biology. 2008 04;6(4):1–16. Available from: https://doi.org/10.1371/journal.pbio.0060102.

Automatic Discovery of Families of Network Generative Processes

Telmo Menezes[1] and Camille Roth[1,2]

{menezes,roth}@cmb.hu-berlin.de

[1] CNRS, Computational Social Science Team at Centre Marc Bloch Berlin (CNRS/HU), Friedrichstr. 191, 10117 Berlin, Germany

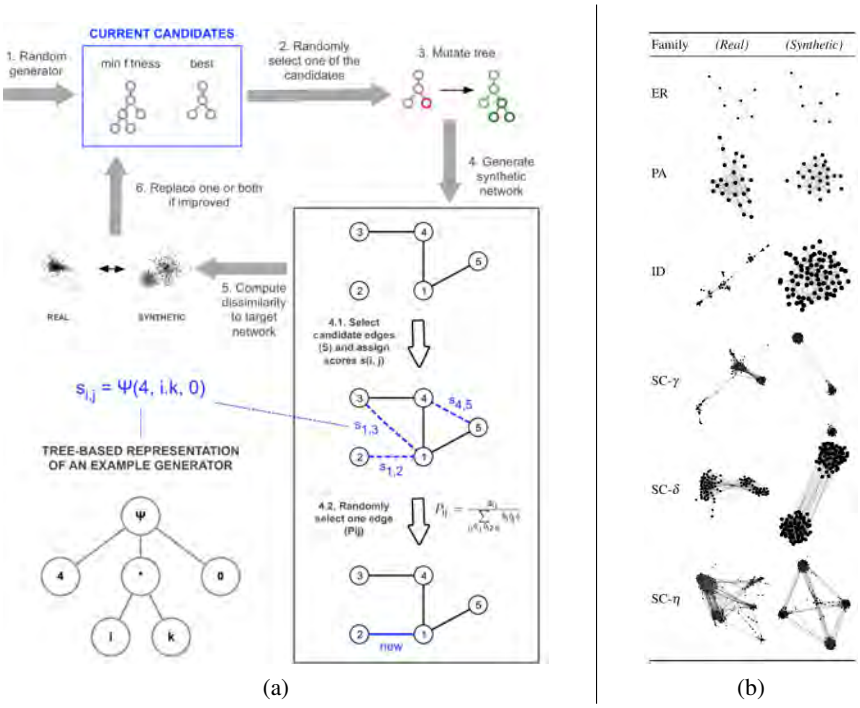
[2] CAMS (Centre Analyse et Mathématique Sociales, UMR 8557 CNRS/EHESS), 54 Bd Raspail, 75007 Paris, France

Note: *This research has recently been published (May 14, 2019) as "Automatic Discovery of Families of Network Generative Processes", by Telmo Menezes and Camille Roth in "Dynamics on and of Complex Networks, Volume III: 'Machine Learning and Statistical Physics' ", and edited by F. Ghanbarnejad, R. S. Roy, F. Karimi, J.-C. Delvenne, B. Mitra, Springer Proceedings in Complexity, pp. 83–111.*

Designing plausible network models typically requires scholars to form *a priori* intuitions on the key drivers of network formation. Oftentimes, these intuitions are supported by the statistical estimation of a selection of network evolution processes which will form the mathematical basis for the development of a stylized model. Machine learning techniques based on evolutionary computation have lately been introduced to assist the automatic discovery of generative models [1,2,3]. Some of these approaches [3] may more broadly be described as "symbolic regression", where fundamental network dynamic functions, rather than just parameters, are evolved through genetic programming. In other words, they aim at automatically discovering plausible *network generation laws* from a given empirical network — i.e. extracting a generative genotype based on a static phenotype.

The core of the present contribution consists in applying symbolic regression to a collection of social networks of the same nature in order to explore the existence of families of regular generative principles for networks of the same realm. In other words, instead of looking for classes of network phenotypes, as is classical in the literature [4], we use symbolic regression to find families of network genotypes, construed as network generators. Our empirical case is based on an original data set of 238 anonymized ego-centered networks of Facebook friends which were randomly sampled from about 10,000 such networks collected in a large-scale online survey.

Figure 1 shows an overview of the generator search process that we employed. Generators are represented as tree-based computer programs, which are equivalent to mathematical expressions. Tree leaves are variables and constants, and its other nodes are operators. For a given candidate pair of nodes, a generator can compute a value that maps to the probability of an arc/edge being created between the nodes. Generators define decentralized growth processes and rely exclusively on local variables expressing topological features of the nodes, such as their current degrees and network distances, as well as unique identifiers. The quality of a generator is evaluated by comparing a synthetic network generated by it with the target network. To measure the similarity



(a) (b)

Figure 1: *(a):* Evolutionary loop including the synthetic network generation process. The outer part of this figure describes evolution at the generator population level, while the framed part on the right describes the evolution of a network for a given generator. *(b):* Visual representation of ego-networks (real) with their reconstruction (left), for a selection of automatically discovered generative families: ER (Erdos-Renyi), PA (preferential attachment), ID-based attachment, and SC (featuring the endogeneous emergence of functions providing for Social Circles)

between the generated (evolved) network and the empirical (target) one, we combine distributions that describe simple aspects of the network, such as in- and out- degree and measures of centrality, with distributions describing finer and more meso-level aspects of the structure, such as distances and triadic profiles. A bias favoring shorter generators is used to avoid overfitting, and to encourage simpler, understandable expressions.

Applying an evolutionary search on each of the above-mentioned Facebook ego-centered networks, we obtain one most plausible per network, thus 238 in total. Comparing generators as mathematical formulas is not a trivial task, but we define an additional measure of similarity, this time between the generator behaviors, in terms of the similarity of the networks that they produce. We then used this measure to produce a two-dimensional embedding of all 238 generators, as shown in figure 2. With the help of this embedding we made easier the task of manual analyzing generators. In particular, we look for patterns of similar generators in mathematical terms, i.e. at the level of the explicit formula.

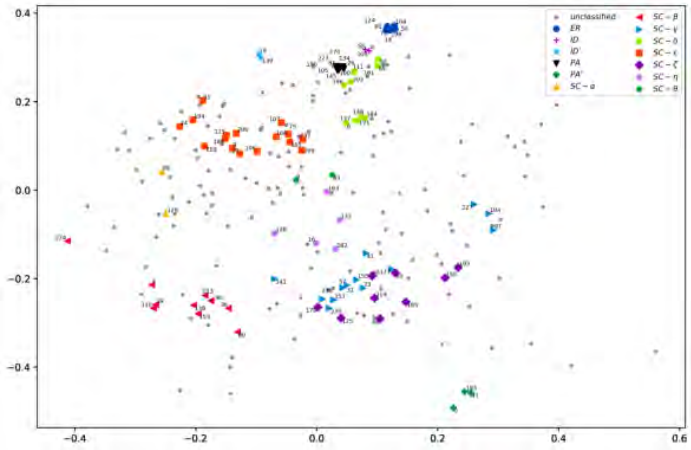


Figure 2: Network generators mapped into a two-dimensional layout according to their pairwise distances. Different colors and shapes indicate families of generators that were manually identified as semantically similar. The legend shows the pattern that identifies each family.

We identified 11 such strong patterns, that we refer to as "*families of generators*". Five of these families are very simple, some matching well-known models such as *preferential attachment* and Erdős-Rényi, especially for smaller networks (which may contain less information or be more basic). The other eight have a very strong resemblance with one another: their link dynamics is strongly influenced by the existence of a certain number of classes of nodes which likely matches underlying *social circles*, i.e. cohesive clusters of nodes. This further yields insights on ego-centered sociability networks, especially with respect to the existence and contribution of social circles in their formation. A simple interpretation for this is indeed that ego networks are a sample of social groups that ego belongs to. For example: school friends, family, work colleagues and so on. It makes sense that these groups are much more densely connected within themselves than between them, as they correspond to separate social spheres.

More broadly, this approach substantiates the existence of a small class of generative behaviors which are widespread among ego-centered networks. It also opens up to the possibility of applying this approach to non-social networks as well.

References

1. Viplove Arora and Mario Ventresca. Action-based modeling of complex networks. *Scientific Reports*, 7(6673), 2017.
2. Kyle Robert Harrison, Mario Ventresca, and Beatrice M Ombuki-Berman. Investigating fitness measures for the automatic construction of graph models. In *European Conference on the Applications of Evolutionary Computation*, pages 189–200. Springer, 2015.
3. Telmo Menezes and Camille Roth. Symbolic regression of generative network models. *Scientific Reports*, 4(6284), 2014.
4. Jukka-Pekka Onnela, Daniel J. Fenn, Stephen Reid, Mason A. Porter, Peter J. Mucha, Mark D. Fricker and Nick S. Jones Taxonomies of networks from community structure. *Physical Review E*, 86(036104), 2012.

Part VIII

Mobility

Scaling behaviours of mobility patterns of e-commerce users

Yuansheng Lin[1,2], Weiran Cai[3], Qianchuang Zhao[1], Yuanqing Wu[2], and Raissa D'Souza[3,4,5]

[1] Department of Automation, Tsinghua University, Beijing, China, 100084,
[2] Beijing Jingdong Century Trading Co., Ltd., Beijing, China, 100101,
[3] Department of Computer Science, University of California, Davis, California, USA, 95616,
[4] Department of Mechanical and Aerospace Engineering, University of California, Davis, California, USA, 95616,
[5] Santa Fe Institute, Santa Fe, New Mexico, USA, 87501

1 Introduction

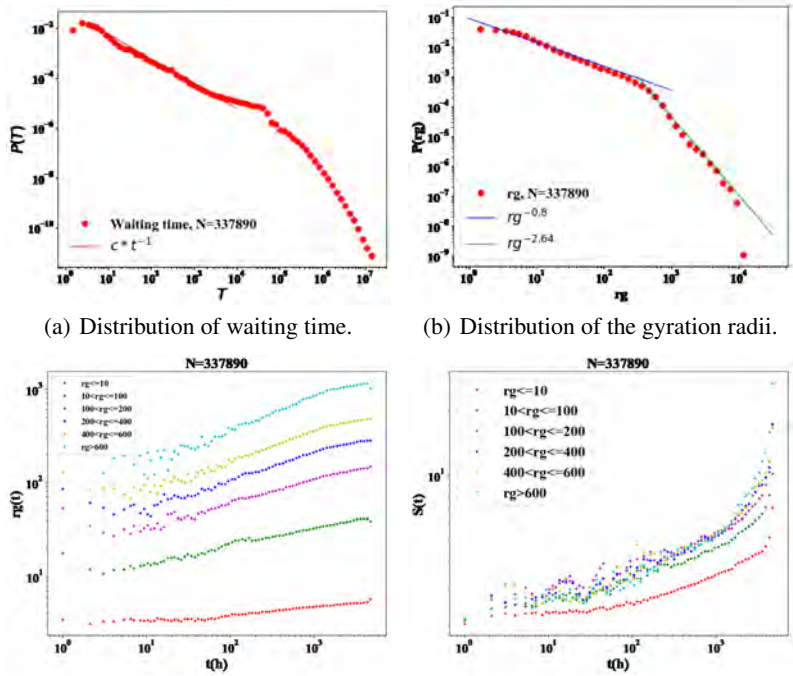
Human behaviours often have associated mobility patterns which can be ubiquitously observed. An in-depth understanding of the laws of human movement would be of great significance in the fields of public health, urban planning and economic forecasting [1–3]. Over the past few years, the availability of data sets, such as dollar-bill tracking and traces of mobile phones, have offered deeper insights for the understanding of human mobility. However, mobility patterns can be distinct for distinct types of human activities. In this work, we collect GPS data for an online-shopping app that has more than 320 million active users. The location data is collected with the users' permission and is only recorded when users are active on the app browsing information on goods, launching an online shopping cart and placing shopping orders. The data set consists of the locations with longitude and latitude and their time of occurrence for the whole year (2018). We analyse the mobility patterns of verified users and find scaling behaviour for the radius of gyration that is different from any previous work [1–3]. A major distinction is that there are two distinct regimes of users, each one obeying a distinct scaling relation. This suggests that understanding of the underlying mechanism that correlates mobility and shopping behaviour is needed. Furthermore, a striking difference also appears in the scaling laws between verified users and those identified as fraud users which enables us to develop a classification algorithm based on XGBOOST to identify fraud users with a high accuracy. Typical fraud behaviours include promo abuse, user abuse or user takeover on the e-commerce platform.

2 Results

We collect location data from a widely-used online-shopping app with the users' permission, recording the users' longitude and latitude, and timestamps, for the whole year (2018). The data set is distinguished from other existing ones for its huge amount of samples and broad coverage of time span and locations, which will reveal rich details in the scaling behaviours.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



(a) Distribution of waiting time.

(b) Distribution of the gyration radii.

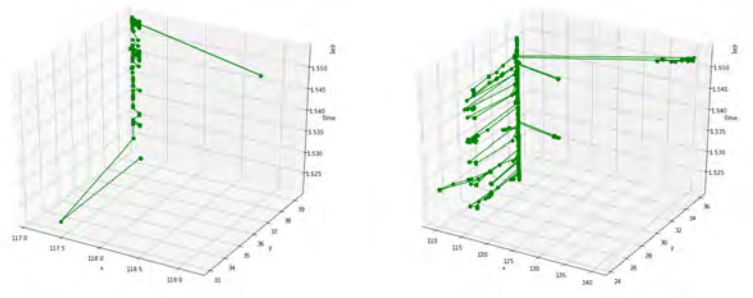
(c) Radius of gyration $rg(t)$ versus time for different rg groups.

(d) The number of visited distinct locations $S(t)$ versus time in a log-log plot.

Fig. 1. Mobility results for normal e-commerce users.

We collect the data of 337,890 users in 2018 who are identified to have used the e-commerce platform in a normal and non-abusing way. We measure the most important metrics that are commonly used to characterize large-scale human mobility patterns. As shown in Fig. 1a, the waiting time distribution $P(T) \propto T^{-1}$ is consistent with the queuing model prediction [1, 4]. However, as shown in Fig. 1b, we find that the scaling exponent for the distribution of radii of gyration $P(rg) \propto rg^{-\alpha}$ is distinct from all previous results. We find that there are two-piecewise scaling modes for population groups with different radii of gyration rg (the population can be divided into two groups, corresponding to $rg < rg^c$ and $rg > rg^c$). The scaling modes are separated by a turning point rg^c, which shows that the behaviour of the population with $rg < rg^c$ is dramatically different from that of the population with $rg \geq rg^c$. Even the scaling before the turning point exhibits a scaling exponent of $\alpha = 0.8$, which is distinct from the previously observed (for example, compared with $\alpha = 1.65$ in Gonzalez, et al [3]).

Moreover, two other scaling behaviours also exhibit peculiarities as shown in Fig. 1c and Fig. 1d. Both the change of users' radii of gyration over time $rg(t)$ and the number of visited locations versus time $S(t)$ are found to be dependent on rg. Such dependence has not been identified in previous work. Two examples of location trajectories are displayed for users with two typical rg values in Fig. 2. The piecewise scaling behavior



(a) *rg*=78. (b) *rg*=602.

Fig. 2. Examples of location trajectories for two normal e-commerce users. Here the x-axis is for longitude, the y-axis is for latitude and the third axis is for the corresponding timestamps.

observed for *rg* is shown in Fig. 1b and the *rg*-dependent behaviour suggest a different mechanism underlying mobility patterns that is associated with the online shopping behaviour, when compared to the pure human movement patterns studied previously. It suggests that the mechanism here is not only affected by the generic mechanisms, exploration and preferential return [1], but also correlated with users' shopping behaviour.

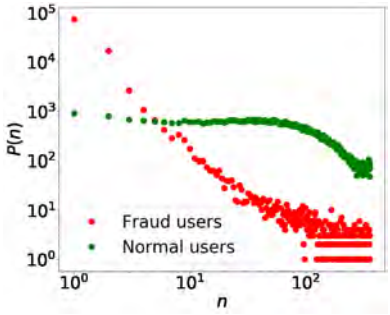


Fig. 3. Distribution of the number of days for which locations are reported for fraud and normal users.

The necessity of examining the system by grouping together similar *rg* users is further demonstrated in our study by classifying the normal and fraud accounts. Fraud users are typically involved in promo abuse, user abuse or user takeover on the e-commerce platform. Astonishingly, we found that adapting to the abnormal shopping behaviours, the mobility patterns of malicious users are clearly distinguishable from those of the normal users (Fig. 3). In particular, the distribution of the number of days for which locations are reported (if a user reports locations everyday in 2018, its number of days is 365) follows completely different scaling rules between the normal and fraud users. Here we have used the Extreme Gradient Boosting algorithm (XGBOOST) for

the classification task [5]. XGBOOST provides parallelization and high predictive accuracy [7, 8]. After removing redundancy and irrelevant information, 24 features including number of active days (Fig. 3) and radii of gyration are found to be more effective in discriminating the two user types with high accuracy (96.78%).

In summary, we have found that the human mobility patterns associated with online-shopping behaviours are different from those shown in pure movements. In particular, the scalings are highly dependent on different classes of users characterized by their radii of gyration: although most metrics still follow power laws, they are separable in terms of the gyration radii. An underlying mechanism taking into account the shopping behaviours is needed to explain the emergence of this new universality class. Moreover, we have shown that the investigation into detailed population structure, with the aid of efficient classification algorithms, may also help to control risks related to underground industries.

References

1. Song C , Koren T , Wang P , et al. Modelling the scaling properties of human mobility. Nature Physics, 2010, 6(10):818-823.
2. Pappalardo L , Simini F , Rinzivillo S , et al. Returners and explorers dichotomy in human mobility. Nature Communications, 2015, 6:8166.
3. Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding individual human mobility patterns. nature, 2008, 453(7196): 779.
4. Barabasi A. The origin of bursts and heavy tails in human dynamics. Nature, 2005, 435(7039): 207-211.
5. Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System. 2016.
6. Sheridan R P, Wang W M, Liaw A, et al. Extreme gradient boosting as a method for quantitative structureactivity relationships. Journal of Chemical Information and Modeling, 2016, 56(12): 2353-2360.
7. Chen X, Huang L, Xie D, et al. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. Cell Death Disease, 2018, 9(1): 3.
8. Wang H, Liu C, Deng L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. Scientific Reports, 2018, 8(1): 14285.

Social influence with recurrent mobility and multiple options. Preliminary results on Swedish data

Attila Szilva[1] and Jérôme Michaud[2]

[1] Department of Physics and Astronomy, University of Uppsala, 752 37 Uppsala, Sweden
attila.szilva@physics.uu.se
[2] Department of Sociology, University of Uppsala, 751 20 Uppsala, Sweden
jerome.michaud@soc.uu.se

1 Introduction

The classic Voter Model (VM) is an extremely idealized description for the evolution of opinions in a population. It has played a central role in probability theory and in statistical physics because it is one of the few exactly solvable many-body interaction systems [1]. Furthermore, many phenomenology rich reality inspired generalizations of the VM have been developed [2]. However, their lack of calibration of the model parameters to *real social data* make them unfit to quantify observations or make predictions.A few steps have been made to overcome this issue [3, 4]. In [3], the Social Influence with Recurrent Mobility (SIRM) has been developed for two-party systems. This generalization of the VM works well for the case when the support of the two parties are well balanced. The key idea of the SIRM model is that the commuting pattern of individuals is a good proxy for the topology of interactions, since people can interact both in the regions where they live and where they work.

While the SIRM model has been shown to successfully recover spatial correlations in U.S. presidential elections [3], it suffers from some mathematical issues in the handling of the stochasticity of the interactions. In this paper, we present a generalized version of the SIRM model that fixes the mathematical issues of the original formulation and that is applicable to multi-party systems [4], as well as a novel calibration procedure for the model parameters and apply it to Swedish data. In addition, we investigate Sweden electoral geography based on Network Science [5] with an analysis that provides a partition of administrative units into *electoral clusters* based on the similarity in their inhabitants' electoral behavior. We present a functional network analysis to uncover stable electoral clusters over time.

2 Method

We now shortly present the SIRM model. Let us denote the fraction of the total population living in region i and working in region j, defining a *commuting cell* ij, by n_{ij} and the vote-share for opinion k in a commuting cell ij by v_{ij}^{k}. The dynamics of the system is controlled by the transition operator $R_{ij}^{kk'} = n_{ij} v_{ij}^{k} p_{ij}^{k\to k'}$. The first factor of the RHS is the probability to choose the communting cell ij, the second factor is the local

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

vote-share for opinion k and $p_{ij}^{k\to k'}$ stands for the probability to change from opinion k to k' and is given [4] by

$$p_{ij}^{k\to k'} = \lambda\left(\alpha v_i^{k'} + (1-\alpha)v'^{k'}_j\right) + \beta\widetilde{v}_{ij,D}^{k'} + \gamma/K\,, \tag{1}$$

where $v_i^{k'}$ and $v'^{k'}_j$ are the vote shares of opinion k' for the population living in region i and for the population working in region j, respectively. The first term encodes recurrent mobility and the parameter α controls the relative importance of interactions at home and at work. The second term controls the noise in the interaction as defined in [4], $\widetilde{v}_{ij,D} = \mathscr{D}ir(v_{ij}/D)$ is a Dirichlet sample of parameter v_{ij}/D and D controls the amplitude of the noise. The third term, called the *free will* term, encodes unilateral change of opinion between the K possible opinions. Finally, parameters λ, β and γ control the relative importance of these three terms. Model parameters D and γ are then calibrated to data according to the procedure developed in [4].

3 Results

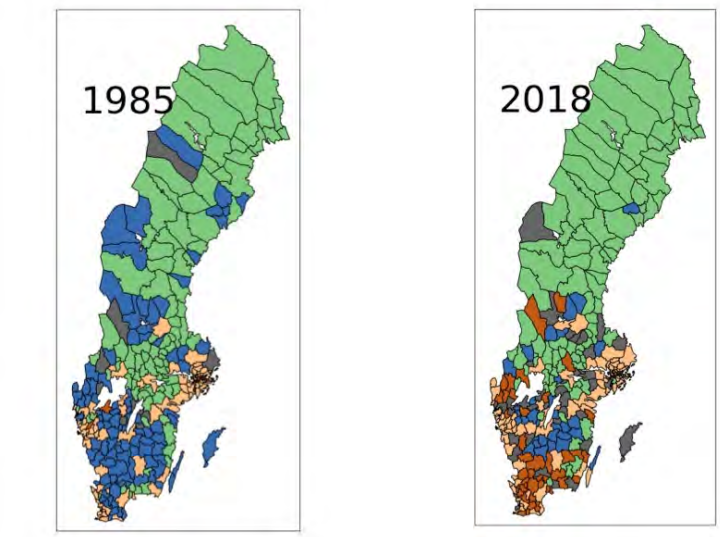


Fig. 1. Electoral clusters in Sweden in 1985 and 2018. The colors are arbitrary and used to visualize clusters identified as the same cluster over time. Municipalities displayed in grey belong to smaller clusters. In the 2018 map, the brown electoral cluster is characterized by a strong vote share for Sweden Democrats.

We start with the results of the functional network analysis [5] applied to the ten parliamentary elections held in Sweden between 1985 and 2018. For each elections, Sweden was partitioned into electoral clusters and these clusters have been shown to

be stable over time. Figure 1 displays the first and last partitions. Three main clusters are present in all election and a fourth emerges and propagates form 2002 (colored in brown in the figure) that is characterize with a stronger than average vote-share for Sweden Democrats (a ring-wing populist party).

The generalized formulation of the SIRM [4] outlined above has been shown to recover the original formulation of the SIRM model [3] in some limit allowing the two formulations to be compared. In the model, we have two different types of noise. The first one models local variations and fluctuations of the probability to change opinion. Here, the challenge was to calibrate of the magnitude of the diffusion constant (D). The other noise, the "free will term" (third term in (1)) allows spontaneous opinion changes which leads to the emergence of new parties in the model.

The calibration procedure developed in [4] has been shown to be robust against coalition, i.e., one can group parties together without changing the calibrated D. This procedure has been applied to a synthetic network and to the commuting network of the Stockholm county. We have discovered that the diffusion constant calculated in the lack of the free will noise term (third term in (1)) for the synthetic case is about 18 times larger than the calibrated value for the U.S. presidential election case calculated in [3], whereas the corresponding value for the Stockholm case is about six times larger. This indicates that a calibration procedure based on stationarity of given statistical properties requires more noise when initial conditions are random than for real initial conditions and that the Stockholm case is more noisy than the U.S. case.

The paper hints to many other possible developments of this work, since the rates can be modified in many ways. One could, for example, use the Dirichlet distribution to add noise on other components of the rate and include time dependent or spatially dependent rates to account for varying and heterogeneous socioeconomic factors that might have an influence on the dynamics of the system. Furthermore, the influence of the commuting network can be studied through numerical experiments.

Summary: In this paper, we discuss the possible generalizations of the Social Influence with Recurrent Mobility (SIRM) model [4]. As a result of the extension, we show that our model works well for multiparty systems and is mathematically well-posed even in the case of extreme vote shares by handling the noise term in a novel way. In addition, we summarize preliminary results of a functional network analysis [5] to the last ten parliamentary elections held in Sweden.

References

1. Holley, R., A., Liggett, T. M. (1975) Ann. Probab. **3**, 643
2. Redner, S., (2019) Reality Inspired Voter Models: A Mini-Review, C. R. Physique
3. Fernández-Gracia, J., Suchecki, K., Ramasco, J.,J., Miguel, M., S., Eguíluz V., M. (2014) Is the Voter Model a Model for Voters? Phys. Rev. Lett. 112, 158701
4. Michaud, J., Szilva, A. (2018) Social influence with recurrent mobility and multiple options, Phys. Rev. E 97, 062313
5. Latora, V., Nicosia, V., Russo, G. (2017). Complex Networks: Principles, Methods and Applications. Cambridge: Cambridge University Press.

Network analysis of internal migration in Austria

Dino Pitoski[1], Thomas Lampoltshammer, and Peter Parycek

[1] Danube University Krems, Krems 3500, Austria,
dino.pitoski@donau-uni.ac.at,
WWW home page: http://www.donau-uni.ac.at/en

1 Introduction

Migration has become an all-important topic in today's political and public discourse. Perhaps in line with that public interest, we have recently observed a surge of publications in which network analysis was applied to the phenomenon of human migration. In the last couple of years alone, more than a dozen of migration-as-network analyses have emerged ([1], [2], to just name a few). Very few network analyses, however, were applied to migration at the level of settlements (cities, towns, and villages). Yet, as the UN forecasts, the rise in urban population will reach about 70% of the total in just a couple of decades. Therefore, we are likely to start discussing the migration at the city level more often than at the country level. Besides that, the data on international (inter-country) migrations are not exact, but only estimated, and, as such, are not useful for the more focused, onspot policy decision making. It is, therefore, imperative to refine the most functional (network) analysis tools for explaining migration within and between cities, or, to be more precise, within and between settlements.

In this vein, due to the lack of such attempts in the past, we have run a network analysis on migration at the settlement level, in the case of Austria. Here we present an excerpt from the network analysis of internal migration in Austria in 2018, with some of the basic network indicators and their comparisons, and an estimation of a gravity model.

2 Network data and definition

Statistics Office Austria defines internal migration in any given year as individuals' changes of address in that same year, recorded at the level of municipalities ([4]). Address changes, independent of the length of stay at any given address, count as migration as long as there is a minimum stay of 90 days in the country as a whole.

Formally, we define the network of Austrian internal migration at any particular year as a weighted directed graph $\mathscr{G} = (\mathscr{N}, \mathscr{L}, \mathscr{W})$, whose:

- nodes $\mathscr{N} = \{n_1, n_2, ..., n_N\}$ represent all Austrian municipalities ($N = 2096$),

- link weights $\mathscr{W} = \{w_{ij}\}_{N \times N}$, are the total counts of recorded residence address changes between or within municipalities occurring within the particular year,

- links $\mathscr{L} = \{l_{ij}\}_{N \times N}$, is a binary projection of $\mathscr{W}$, such that $l_{ij} = 1$ if $w_{ij} > 0$ and $l_{ij} = 0$ if $w_{ij} = 0$.

In this general formulation, we take into account loops ($w_{ii} \geq 0$). From $\mathscr{G}$ we further identify a (spanning) subgraph $\mathscr{G}' = (\mathscr{N}, \mathscr{L}', \mathscr{W}')$, where $\mathscr{W}' = \mathscr{W} \setminus \{w_{ii}\}$ and $\mathscr{L}'$ is the according binary projection of $\mathscr{W}'$.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

3 Results

We observe that the network of Austria-internal migration in 2018 (Figure 1) is dominated by loops.

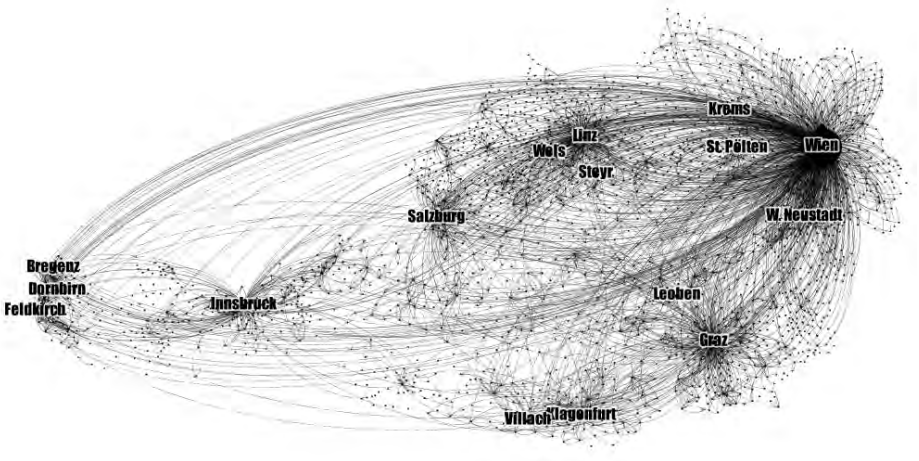


Fig. 1. Austria-internal migration network 2018. Labels indicate the most central nodes in $\mathscr{G}$ in terms of their total node strength (s_i). With little variation in ranking, the same nodes appear as highest-strength nodes in $\mathscr{G}'$. Edges of $w_{ij}, wii \leq 10$ have been omitted from visualization.

We find very high correlation between direction-respective node strength values ([5]) for $\mathscr{G}$ vs. for $\mathscr{G}'$; $\rho(s_i^{in}, s_i^{in\prime}) \approx 0.96$, $\rho(s_i^{out}, s_i^{out\prime}) \approx 0.97$. We also find nearly perfect correlation ($\rho > 0.98$) between in- and out- degrees of nodes within both $\mathscr{G}$ and in $\mathscr{G}'$and both in weighted and projected binary view. In that regard, in $\mathscr{G}'$, we search for and find, expectedly, very high weighted reciprocity ([6]), $r' = \frac{\sum_i \sum_{j \neq i} w_{ij}^{\leftrightarrow\prime}}{\sum_i \sum_{j \neq i} w_{ij}\prime} \approx 0.81$ (≈ 0.47 in binary network), while the average value of the non-reciprocated weights (ibid.) is $w_{ij}^{\overleftarrow{\leftrightarrow}\prime} \approx 4$ ($\max w_{ij}^{\leftrightarrow\prime} = 497$).

We further test the hold of the gravity law ([7]) on $\mathscr{G}'$. We produce estimated weights as $\hat{w_{ij}}' = P_i P_j {d_{ij}}^{-1}$, where P_i and P_j are populations of origin and destination in 2018, respectively (data obtained from Statistics Office Austria), and d_{ij} are the, currently, shortest driving distances between them (data obtained using Google's Distance Matrix service). A simple regression shows very close fit between real and estimated weights, and most of the migrations occurring among places separated by shorter (< 100 km) driving distances (Figure 2).

Summary. We provide a rough sketch of the nature of migration occurring within Austria by using network indicators of centrality, reciprocity and through the gravity model. Our results suggest Austria-internal migration takes place mostly within the boundaries of large cities (or periphery-city), else mostly between larger cities, and that these larger cities tend to send/receive migrants to/from many diverse locations. We find high reciprocity and symmetry of migration flows; for each specific migration link A to B, there

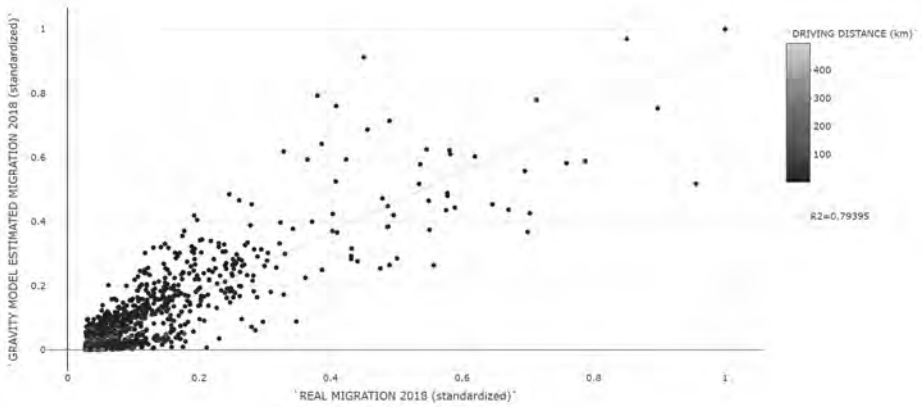


Fig. 2. Gravity law in Austria-internal migration 2018. Linear regression on real vs. gravity law-estimated values, standardized as $w'_{ij}/\max w'_{ij}$, $\hat{w_{ij}}'/\max \hat{w_{ij}}'$. The model is reduced to the 1300 links of highest weights in $\mathscr{G}'$.

are roughly same-sized countermigrations B to A. The results also show that the base gravity law variables, of settlement populations sizes and driving distances between settlements serve as very good predictors of internal migration in Austria. Our ongoing research is invested into testing the greater variety of network indicators, both those existing, as well as customized for the specific case, and developmental models, especially the updates and modifications to the gravity law model.

Acknowledgements. This study was co-financed by the Asylum, Migration and Integration Fund and the Austrian Federal Ministry of the Interior, Project No. 141235506.

References

1. Danchev, V., Porter, M., A.: Neither global nor local: Heterogeneous connectivity in spatial network structures of world migration. Social Networks, 53, 419. (2018)
2. Charyyev, B., Gunes, M., H. (2019). Complex network of United States migration. Computational Social Networks 6, 1.
3. United Nations.: World Urbanization Prospects: The 2018 Revision: key facts. Available at https://population.un.org/wup/Publications/Files/WUP2018-KeyFacts.pdf, accessed 26th Aug 2019. (2018)
4. Bundesanstalt Statistik Österreich: Standard-Dokumentation Metainformationen (Definitionen, Erluterungen, Methoden, Qualitt) zur Wanderungsstatistik. Available at www.statistik.at/web_de/wcmsprod/groups/gd/documents/stddok/029352.pdf, accessed 29th Aug 2019. (2014)
5. Barrat, A. and Barthélemy, M. and Pastor-Satorras, R. and Vespignani, A.: The architecture of complex weighted networks. Proceedings of the National Academy of Sciences, 101(11), 37473752. (2004)
6. Squartini, T., Picciolo, F., Ruzzenenti, F., Garlaschelli, D.: Reciprocity of weighted networks. Scientific Reports, 3(1). (2013)
7. Zipf, G. K.: The P 1 P 2 D Hypothesis: On the Intercity Movement of Persons. American Sociological Review, 11(6), 677. (1946)

Analyzing patterns of mobility and internal migration among researchers in Mexico using longitudinal bibliometric data

Andrea Miranda González[1,2], Samin Aref[2],
Tom Theile[2], and Emilio Zagheni[2]

[1] Department of Demography, University of California Berkeley, Berkeley, CA, USA
andrea.mirgon@berkeley.edu
[2] Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, Rostock, Germany
{aref,theile,zagheni}@demogr.mpg.de

1 Introduction

The academic exchange of ideas can go beyond physical borders. As such many scholars are highly mobile and their work contributes to their host, rather than their origin countries, through technological and economic advances. A growing body of literature focuses on the mobility of scientists and its impact at the international level. From the public policy perspective, it is in the interest of countries to maintain a strong base of highly qualified scholars who can provide innovative and scientific solutions to public issues. In doing so, governments look for the underlying reasons for researchers' movements and sources of attraction at national and global level.

Nevertheless, little is known about the drivers of internal migration of researchers. Understanding these patterns can shed light on important regional deficits that are the source and outcome of disparities and inequality of opportunities for future generations. We propose an approach to study internal migration of scholars using Scopus bibliometric data. We present our methods to measure mobility within Mexico as well as interpret it from a network perspective. Mexico is a particularly important case for exploratory analysis because a larger share of its mobile population moves internally rather than internationally. Between 2005 and 2010, interstate and intrastate mobility represented 3.5% and 3.1% relative to 1.1% of the population moving abroad [11]. Moreover by focusing on Mexico, we study an emerging system of science which has several leading universities of Latin America. In addition, Mexico is an under-studied case in scientometrics literature and it remains unclear whether mobility in Mexico has increased or slowed down as a result of special socioeconomic conditions, such as government spending on public institutions, social inequality, and alternative jobs in the private sector. This analysis intends to contribute twofold to the literature: first, by re-purposing bibliometric data to analyze internal rather than international migration, secondly by exploring mobility patterns of scholars in Mexico. Although our substantive focus is on Mexico, the proposed methodological framework of re-purposing bibliometric data for internal migration is applicable to other countries.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Data and Methodology

For analyzing mobility of researchers, many studies have relied on bibliometric databases such as Scopus [9,8]. Compared to other bibliometric databases, Scopus provides a wider breadth of records in varied disciplines [6] and offers a more reliable author ID [7] which is suitable for tracking mobility of individual researchers [1]. Other recent studies offer proxies for place of residence [4], provide bilateral international migration flows [5], offer a methodological framework for dealing with multiple affiliations [10], and analyze mobility of highly mobile researchers and return migration [2].

Large-scale bibliometric data allow us to identify movements of researchers in a way which has not been possible with traditional sources of migration data like censuses and surveys. The unit of the data is *authorship record* which is the linkage between an author and a publication. Our data involve 1.1 million authorship records of scholars who have published with Mexican affiliation addresses in sources covered by Scopus. Using the data, we analyze mobility patterns of over 200,000 researchers between 32 states of Mexico through the changes in their affiliation addresses over the 1996-2016 period. Prior to the analysis, the data were pre-processed in order to extract the state of the institution of affiliation for each scholar in a given year. First, a state-detection algorithm is used to identify the most likely state from different parts of a given authorship record. Then the results, combined with manually extracted states for 2200 records, were used as training data for developing a neural network using *Keras* [3] which identifies the state for a given authorship record with an accuracy of 98.9%.

3 Results and Discussion

During the period 1996-2016, the majority of scholars have remained in one state and only 7.8% have moved between states. The data show that the median mobile scholar has actively published for 9 years while its non-mobile counterpart only for 5 years.

Although Mexico City appears to attract many scholars, the consistent and negative net migration rate in Figure 1 suggests that more scholars have exited than entered. However Jalisco, an important economic actor of the Pacific coast region, is an example of a common trend in other states where migration rates vary greatly.

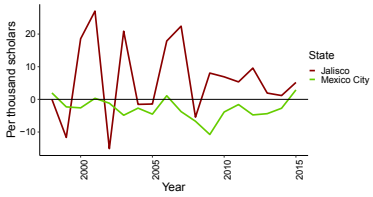
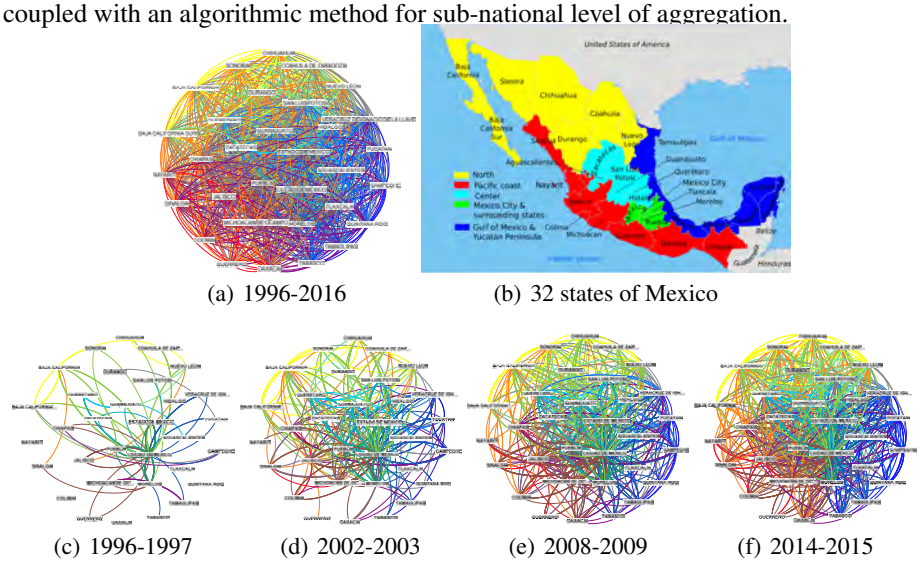


Fig. 1. Net migration rates for scholars in selected states

Figure 2 shows the direction and magnitude of movements of scholars in Mexico between 1996 and 2016. The states that receive and emit the most scholars include the capital city and its surrounding states (State of Mexico, Puebla, and Morelos), as well as states that contribute the most to national GDP such as Nuevo Leon, Guanajuato, Jalisco and Michoacan. Overall, Mexico City appears to be the main destination and origin of mobile scholars, which may be due to its political and economic importance

as well as housing several large national universities. Subpanels (c-f) of Figure 2 highlight the period movements of scholars between states. Overall, the mobility network of researchers has not only become more dense but also more diverse over the past two decades. For instance, in more recent years, states along the Pacific coast (red) show a greater exchange (purple edges) with states along the Gulf of Mexico and the Yucatan Peninsula (blue).

By studying the changes in the migration flows and rates of scholars between the 32 Mexican states, we offer a general perspective of where scholars are attracted to move to. We also analyze general traits of scholars such as their number of years of active publication and the main states of origin and destination. Our results suggest that there is heterogeneity in the direction and magnitude of scholarly movements while Mexico City and its surrounding states appear frequently on the paths of mobile researchers based on betweenness centrality measure. Finally, our work highlights that longitudinal bibliometric data offer valuable insight into internal migration patterns of scholars when coupled with an algorithmic method for sub-national level of aggregation.



(a) 1996-2016 (b) 32 states of Mexico

(c) 1996-1997 (d) 2002-2003 (e) 2008-2009 (f) 2014-2015

Fig. 2. Network of internal migration among researchers in Mexico in 1996-2016 (a), a map of the colored regions corresponding the nodes of the networks (b) four cross-sectional networks based on selected one-year periods (c-f). Directions of edges are clock-wise and their colors are the mix of respective origins and destinations. Intensity of movements is seen by the thickness of the edges (see the figure on screen for high resolution).

Acknowledgements

The authors are grateful to Kompetenzzentrum Bibliometrie and Max Planck Digital Library for providing access to the bibliometric data used in this study and to CONACYT for their support. Comments from the anonymous referees and data quality checks by Jakob Voigt were highly appreciated by the authors.

References

1. Aman, V.: Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. Scientometrics 117(2), 705–720 (Nov 2018), https://doi.org/10.1007/s11192-018-2895-3
2. Aref, S., Zagheni, E., West, J.: The demography of the peripatetic researcher: Evidence on highly mobile scholars from the web of science. Lecture Notes in Computer Science (2019), proceedings of the 11th International Conference on Social Informatics
3. Chollet, F., et al.: Keras: The Python deep learning library. https://keras.io (2015)
4. Czaika, M.: High-skilled Migration: Drivers and Policies. Oxford University Press, New York, NY, USA (2018)
5. Czaika, M., Orazbayev, S.: The globalisation of scientific mobility, 1970–2014. Applied Geography 96, 1–10 (2018)
6. Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G.: Comparison of PubMed, Scopus, Web of Science, and Google scholar: Strengths and weaknesses. The FASEB journal 22(2), 338–342 (2008)
7. Kawashima, H., Tomizawa, H.: Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. Scientometrics 103(3), 1061–1071 (2015)
8. Moed, H.F., Halevi, G.: A bibliometric approach to tracking international scientific migration. Scientometrics 101(3), 1987–2001 (2014)
9. Moed, H.F., Plume, A., et al.: Studying scientific migration in Scopus. Scientometrics 94(3), 929–942 (2013)
10. Robinson-García, N., Sugimoto, C.R., Murray, D., Yegros-Yegros, A., Larivière, V., Costas, R.: The many faces of mobility: Using bibliometric data to measure the movement of scientists. Journal of Informetrics 13(1), 50–63 (2019)
11. Téllez Vzquez, Y., López Ramírez, J., Romo Viramontes, R.: Prontuario de migración interna (2014), https://www.gob.mx/cms/uploads/attachment/file/99594/Prontuario_Migracion_Interna_2013.pdf

Part IX

Multilayer Networks

Corpus linguistics and language networks: A new perspective from the concepts of line graph and multilayer network

Angeles Criado-Alonso, Elena Battaner, Miguel Romance and Regino Criado

Rey Juan Carlos University, Tulipán s/n, 28933-Móstoles, Madrid, Spain
angeles.criado@urjc.es, elena.battaner@urjc.es,
miguel.romance@urjc.es, regino.criado@urjc.es

1 Introduction

Even though it is a known fact that many of the great advances of science appear by deepening into topics that are at the frontier of two or more scientific fields, the study of language from the perspective of the theory and tools of complex networks has a certain tradition [7, 11–14, 18]. Analyzing a particular system, discovering a complex network related to it and studying the properties of that network in order to draw conclusions about the system analyzed is a methodology that has produced a large number of results of great interest in many applications [1, 2, 16]. The research on the system under study must necessarily encompass a diversity of views including different complementary aspects of the network structure. Throughout this study, a corpus is a collection of authentic texts collected electronically according to a set of specific criteria used as a representative sample of a language or subset of that language [4]. In our case, we are interested in the study of the mathematical language produced by the scientific community about complex networks. The complex network arisen from this study roots on a linguistic corpus composed by 89 papers and extended abstracts (all of them based on the theory and applications of complex networks) and a total of 147,637 words and 25,210 sentences. This complex network will be used to design a help tool for specialized translations of this scientific area.

A central assumption of modern linguistics is that language is a system [7]. Moreover, it can be said that language is not only a network, but a complex network, which with appropriate research can and should be fully exploited as an efficient and effective approach to linguistic study [7]. Thus, the analysis of linguistic theories based on the study of its corpus and the vision provided by complex networks can reflect stylistic and typological characteristics of languages, contributing significantly to the search for and establishment of the underlying laws and properties of the human being.

As manifested in [7], there is a wide range of quantitative measures [1, 2, 16] available for the characterization of the topological properties of a linguistic network. For our analysis we are interested in considering a bi-layer network [3, 10] and the concept of line graph [8, 9] which is very useful to highlight he importance that edges have sometimes over nodes in the context of some networks and graphs. In fact, the main motivation behind our study is the new vision that provides the concept of line graph to characterize the structure of a language, allowing the realization of a comparative

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

analysis between different grammars, and the clarity and new approach that offers the use of a multilayer model for the analysis of the corresponding network.

At this point it is important to highlight that there are several approaches when analyzing a language from the perspective of complex networks [5–7, 11–15, 17] but all of them are different from the one we are proposing here. In our model we consider a directed network built from the corpus under study: network nodes are the words that appear in any of the texts that make up the corpus, establishing a (directed) link connecting two words if they appear consecutively (co-occurrence directed) somewhere in a text. In addition, we place the nodes in two different layers. One layer, the layer N_1, is formed by the "empty words" (v.g,: the, of, and, a, in, to, for, ...) and the other layer, the layer N_2 is formed by substantives and specific terms of the specialized language under study (v.g.: network, nodes, systems, model, matrix,...). So, we have a directed network $G = (N, E)$ with two layers, in which N is the set of different words of the text ($N = V_1 \cup V_2$) and E is the set of directed edges (intra and interlayer).

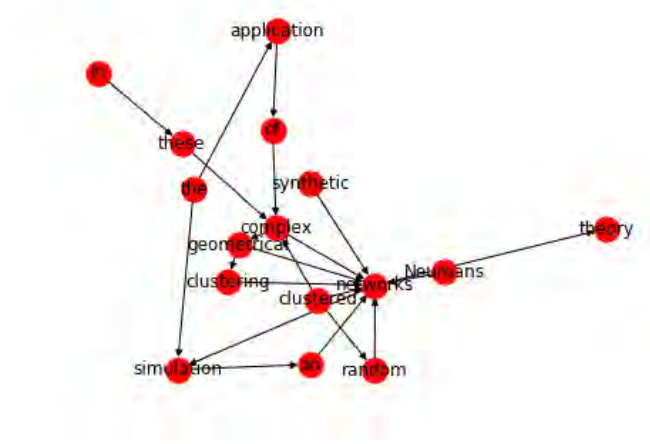


Fig. 1. Network built with the words surrounding the word network in 6 randomly chosen texts

2 Results

All natural languages have syntax, which encodes the relationships between concepts (semantic structures) and underlies the linear sequencing of words [7, 13, 14]. The model we present allows the analysis of the interaction between specific terms of the text through the usual parameters of network theory (degree, average length of paths, ...) in the field of multi-layer networks, being of capital importance the properties of the line graph of (N_1, E_1) (for example, the length of the paths of this layer) that allow to compare texts and to classify them according to the Common European Framework of Reference for Languages (A1,A2,B1,B2,C1,C2). A random text generator based on a random walker on this model is presented. An additional layer with terms equivalent to

those appearing in the corpus will allow to increase the linguistic complexity level of the text. Finally, the model constitutes a very useful tool for translators working with specialized texts in the area of knowledge corresponding to the corpus studied, since a quality translation has to maintain the main linguistic characteristics used in the source text, a especially difficult task if the translator is not mastered in this field. An extension of this line of research will make it possible to compare corpus of different languages according to the topological properties of the corresponding networks.

References

1. Albert, R. and Barabasi, A. L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002).
2. Boccaletti, S., Latora, V. ,Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics, Phys.Rep. 424 75–308 (2006).
3. Boccaletti S., Bianconi G., Criado R., Del Genio C.I., Gómez-Gardeñes J., Romance M., Sendiña-Nadal I., Wang Z., Zanin, M.: The structure and dynamics of multilayer networks. Phys. Rep. 544 (1):1–122 (2014).
4. Bowker, L.,Pearson, J.: Working with Specialized Language: A practical guide to using corpora, Routledge (2002).
5. Cárdenas, J.P., et al.: Topological Complexity in Natural and Formal Languages, Int. J. Complex Systems in Sciences, vol.1(2), 221–225 (2011).
6. Cárdenas, J.P., Olivares, G., Alfaro, R.: Clasificación automática de textos usando redes de palabras, Revista signos: estudios de lingüstica 86, 346-364 (2014).
7. Cong, J., Liu, H. : Approaching human language with complex networks, Physics of Life Reviews 11(4) (2014).
8. Criado, R., Flores, J., García del Amo, A., Romance, M.: Analytical relationships between metric and centrality measures of a network and its dual. JCAM 235 (7): 1775-1780 (2011).
9. Criado, R., Flores, J., García del Amo, A., Romance, M.: Structural properties of the line-graphs associated to directed networks. Networks and Heterogeneous Media 7 (3): 373-384 (2012).
10. De Domenico, M., Solé-Ribalta,A., Cozzo, E., Kivela, M., Moreno, Y., Porter, M.A., Gómez, S., Arenas, A.:Mathematical formulation of multi-layer networks. Phys. Rev. X 3: 041022 (2013).
11. Ferrer i Cancho,R., Solé, R.V.: The Small World of Human Language, Proc. of the Royal Soc. of London B, 286:2261-2266 (2001).
12. Ferrer i Cancho,R., Riordan, O., Bollobás,B.: The consequences of Zipf's law for syntax and symbolic reference, Proc.Biol. Sci/The Royal Society 272 (1562): 561–565, (2005).
13. Liu, H., Hu, F.: What role does syntax play in a language network?. EPL (Europhysics Letters) 83, 18002 (2008).
14. Liu, H., Xu, C., Liang, J.: Dependency distance: a new perspective on syntactic patterns in natural languages. Physics of life reviews 21, 171-193 (2017).
15. Martinčić-Ipšić, S., Margan, D., Meštrović, A.: Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. Physica A 457, 117–128, (2016).
16. Newman, M.: Networks: an introduction. Oxford University Press (2010).
17. Solé, R.: Syntax for free?, Nature 434: 289 (2005).
18. Zipf, G.L.: Human Behavior and the Principle of Least Effort: Hafner (1965).

Analysis of Temporal Change of Japanese Interfirm Transaction Relations as a Multilayer Network

Hitomi Sato[1], Haruka Kato[1], Yuichi Kichikawa[2], Hiroshi Iyetomi[2,5], Ryohei Hisano[3,5], and Tsutomu Watanabe[4,5]

[1] Graduate School of Science and Technology, Niigata University, Niigata 950-2181, Japan,
[2] Faculty of Science, Niigata University, Niigata 950-2181, Japan
[3] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan
[4] Graduate School of Economics, The University of Tokyo, Tokyo 113-0033, Japan
[5] The Canon Institute for Global Studies, Tokyo 100-6511, Japan

1 Introduction

One can describe time evolution of complex systems in biology, computer science, economics, and sociology as multilayer networks [1]. A network representing interactions (links) between basic entities (nodes) is constructed at every time step and the snapshot layers are combined to form a multilayer network by connecting common nodes appearing in the adjacent layers. Here we report an empirical study on temporal change of Japanese interfirm transaction relations taking such a promising approach with an emphasis on their community structure.

2 Construction of a multilayer network

We use annual data on transaction relations between firms in Japan compiled by Teikoku Databank, Ltd. to construct a network for this study. The dataset covers the period of 2003 through 2012, including the Lehman crisis in 2008 and the Great East Japan Earthquake in 2011. The numbers of firms (nodes) and transaction relations (links directed from suppliers to buyers with the same weight) are approximately 125,000 and 730,000, respectively, with variations of less than 10% during the decade; see Refs. [2, 3] for details of the dataset. Here we stress that the network is so dynamic that about 60% of the links which exist in 2003 are replaced by new ones in 2012.

Since we focus on evolution of the community structure due to changes in transaction relations between firms, we first build a link network [4] in each year which is complementary to the original node network; links in the original network correspond to nodes in the link network. The following corresponding relations are established between a node network and its companion:

- The WCC (weakly connected component) of the link network is the WCC of the original node network and the opposite is not always true.
- The SCC (strongly connected component) of the link network is the SCC of the original node network and the opposite is also true.
- The bow-tie structure of the link network is equivalent to the bow-tie structure of the original node network.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

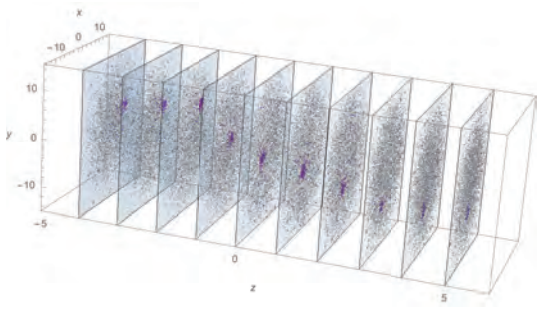


Fig. 1. Evolution of the Japanese interfirm transaction rel
work in three dimensional space. Individual layers ($x-y$
multilayer network are snapshots of the network taken at e
purple color are nodes belonging to the fifth largest com
algorithm for the multilayer network as shown in Fig. 3.

In this study, we analyze the giant SCC of the link network, which encompasses about 80% of the whole system. We then construct a multilayer network by regarding the link networks as layers. To connect common nodes in adjacent layers in both directions, we assume that the relative weight values for interlayer and intralayer links are given by ξ ($0 \leqslant \xi \leqslant 1$) and $1-\xi$, respectively.

Figure 1 visualizes the multilayer network thus constructed in three-dimensional space with $\xi = 0.5$. The nodes are arranged in the x-y plane using a spring-electric model in which linked nodes are attracted by spring force and all nodes are repelled each other by Coulomb force. The z axis represents the time direction. The x-y planes with $z = -4$ and $z = 5$ are layers in 2003 and 2012, respectively. This visualization allows us to see the parts where the original links are dense.

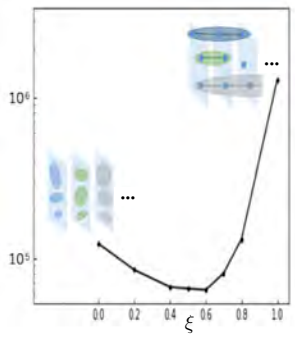


Fig. 2. The number of communities detected by the map equation algorithm for the multilayer network as a function of the interlayer coupling parameter ξ. Note that the y axis is in logarithmic scale.

3 Community Detection

To elucidate temporal change of the Japanese interfirm transaction relations, we illuminate how evolve major communities in the multilayer network by adopting the map equation algorithm [5], a flow-based community detection method. Figure 2 shows the number of communities as a function of the interlayer coupling parameter ξ. When $\xi = 0$, communities in the multilayer network are identical to those in the layers of individual years. Increment of ξ from $\xi = 0$ connects the communities so separated in time direction. When $\xi = 1$, on the other hand, identical nodes (links in the original network) sequentially connected in time direction form communities; the maximum size of those communities is hence 10. Decrement of ξ from $\xi = 1$ joins the communities

within the layers. Therefore, there should be an optimum value of ξ at which the number of communities is the smallest. In fact, the community structure is optimized around $\xi = 0.6$. We note that the community size obeys a power-law distribution.

Figure 3 is an evolutionary diagram for the communities of the multilayer network obtained at $\xi = 0.6$, where only the 10 largest ones are shown. The communities distinguished by different colors are piled up vertically so that larger communities are placed at lower positions. The communities show various types of evolutionary patterns. Some exist steadily over the 10 years, some gradually fade out, some gradually emerge, some appear in the middle of the period. Specifically, the fifth community colored purple is very stable over the 10 years. The dots of the same color in Fig. 1 demonstrates how stable is the community.

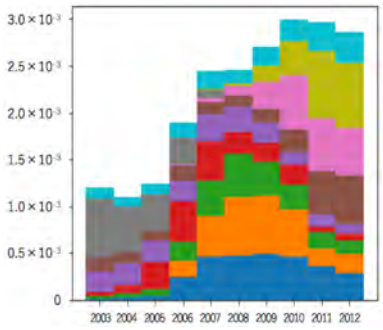


Fig. 3. Evolution diagram of the 10 largest communities in the Japanese interfirm network, where the vertical axis shows the cumulative relative size of the communities.

Summary. We are thus successful in elucidating how the interfirm transaction network in Japan develop gradually with a special emphasis on its community structure. More detailed analyses on the evolution of the community structure are in progress. This work was partially supported by JSPS KAKENHI Grant Numbers 17KT0034, 18K03451.

References

1. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks, Journal of Complex Networks, 2(3), 203–271 (2014)
2. Mizuno, T., Souma, W., Watanabe, T.: Buyer-supplier networks and aggregate volatility. In The Economics of Interfirm Networks, Springer, pp. 15–37 (2015)
3. Hisano, R., Watanabe, T., Mizuno, T., Ohnishi, T., Sornette, D.: The gradual evolution of buyer-seller networks and their role in aggregate fluctuations. Applied Network Science, 2(1), 9 (2017)
4. Luo, J., Magee, C.L.: Detecting evolving patterns of selforganizing networks by flow hierarchy measurement, Complexity 16(6), 53–61 (2011)
5. Bohlin, L., Edler, D., Lancichinetti, A., Rosvall, M.: Community detection and visualization of networks with the map equation framework. In Measuring Scholarly Impact, Springer, pp. 3–34 (2014)

Parametric control of PageRank centrality by using personalization vectors: Classic and Biplex models

Miguel Romance[1,2,3], Regino Criado[1,2,3] Julio Flores[1,2], Esther García[1], and Francisco Pedroche[4]

[1] Departamento de Matemática Aplicada, Ciencia e Ingeniería de Materiales y Tecnología Electrónica, ESCET, Universidad Rey Juan Carlos, C/Tulipán s/n, 28933 Móstoles (Madrid), Spain

[2] Center for Computational Simulation, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón (Madrid), Spain

[3] Data, Complex Networks and Cybersecurity Research Institute, Univ. Rey Juan Carlos, 28028 Madrid, Spain

[4] Institut de Matemàtica Multidisciplinària, Universitat Politècnica de València, 46022 València, Spain

miguel.romance@urjc.es

1 Introduction

Roughly speaking Science tries to explain and understand any phenomenon that occurs in real life. In order to reach this goal, the scientific activity can be classified into three categories: observation, prediction and control. In this presentation we are focusing on the control of the centrality of a complex networks in terms of some parameters of the functions considered. There are many different centrality measures in Networks Science, including local parameter (such as the in-degree), metric parameters (such as the betweenness centrality) and spectral centralities (such as the eigenvector centrality), but the PageRank centrality plays a relevant role, since it has many relevant applications [6]. This measure is the basic ingredient of the (probably) most famous web searcher (Google), but it also has many applications to different real-life problems, ranging from biological systems to cibersecurity (hacking detection).

Given a complex network $G = (X,E)$ of n nodes, a stochastic vector (so called *personalization vector*) $\mathbf{v} \in \mathbb{R}^n$ and $\alpha \in (0,1)$ (so called *dumping factor*), the (classic) PageRank of the network is defined as the steady state of the Markov chain whose transition matrix is given by

$$G = \alpha P + (1-\alpha)\mathbf{e}\mathbf{v}^T,$$

where $\mathbf{e} = (1,1\cdots,1)^T \in \mathbb{R}^n$ and P is the row-stochastic matrix associated to the adjacency matrix $A = (a_{ij})$ of the graph, i.e. if $P = (p_{ij})$, then

$$p_{ij} = \frac{a_{ij}}{k_{out}(i)} = \frac{a_{ij}}{\sum_k a_{ik}}.$$

An alternative way of defining a PageRank-like centrality measure is considering a biplex point of view [8] such that each navigation mode (i.e random walking by using

the connections of the network and random walking by using the personalization vector) corresponds to a layer in a multiplex network [1]. Hence the Biplex PageRank is a n-dimensional projection of the steady state of the Markov chain whose transition matrix is given by

$$M = \begin{pmatrix} \alpha P & (1-\alpha)I \\ \alpha I & (1-\alpha)\mathbf{e}\mathbf{v}^T \end{pmatrix} \in \mathbb{R}^{2n\times 2n}.$$

By using these ingredients, we can study the influence of the parameters of such centrality measures (either their damping factors or their personalization vectors) in the possible values of the PageRank centralities. This controllability analysis was also performed for other spectral centrality measures in [5] and for the damping factor in the case of the (classic) PageRank in [2], so in this presentation we will focus on the influence of the personalization vector in the Classic and Biplex PageRanks [4, 7, 3].

2 Results

Our first analytical results show that we can give a sharp localization of all possible PageRank centralities obtained for all admitted personalization vectors [4, 7].

Theorem 1 ([4], Theorem 3.2). *If we denote by $\mathscr{PR}_\alpha(i)$ the set of all possible values of (personalized) classic PageRank of node $i \in \{1,\ldots,n\}$ and fixed $\alpha \in (0,1)$, then*

$$\mathscr{PR}_\alpha(i) = (\min_j x_{ji}, x_{ii}),$$

where $X = (x_{ij})$ is the matrix given by

$$X = (1-\alpha)\,(I-\alpha P)^{-1}. \tag{1}$$

Theorem 2 ([7], Theorem 3.9). *If we denote by $\mathscr{PRB}_\alpha(i)$ the set of all possible values of the Biplex PageRank of node $i \in \{1,\ldots,n\}$ and fixed $\alpha \in (0,1)$, then*

$$\mathscr{PRB}_\alpha(i) = (\min_j c_{ji}, c_{ii}),$$

where $C = (c_{ij})$ is the matrix given by

$$C = \frac{(1-\alpha)^2}{\beta}\left(I - \frac{\alpha}{\beta}P\right)^{-1}((1+\alpha)I - \alpha P), \tag{2}$$

with $\beta = 1-\alpha(1-\alpha)$.

We can also analyze the relative position of the intervals $\mathscr{PR}_\alpha(i)$ and $\mathscr{PRB}_\alpha(i)$, obtaining the following analytical result:

Theorem 3 ([3], Theorem 3.1). *Let $G = (X,E)$ be a complex network with n nodes and no loops. If $i \in \{1,\ldots,n\}$, then $\mathscr{PR}_\alpha(i) \cap \mathscr{PRB}_\alpha(i) \neq \emptyset$ for all $\alpha \in (0,1)$.*

Since $\mathscr{PR}_\alpha(i) \cap \mathscr{PRB}_\alpha(i) \neq \emptyset$, we can compute all possible relative positions of such intervals for different families of random networks (see [3]). In particular, if we consider Barabási-Albert synthetic networks with $n = 100$ nodes and compute the relative position of intervals for different values of the minimum degree value p from 5 to 40, Figure 1 shows that for small values of α all nodes verifies that $\mathscr{PR}_\alpha(i) \subseteq \mathscr{PRB}_\alpha(i)$ (right panel). On the other hand, if $\alpha \geq 0.5$ (independently of the minimum degree value d) all nodes verifies that $\mathscr{PRB}_\alpha(i) \subset \mathscr{PR}_\alpha(i)$ (left panel).

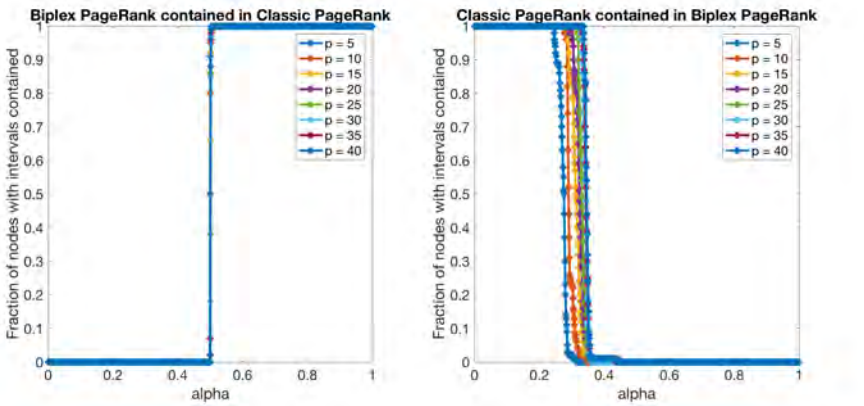


Fig. 1. Relative position of intervals $\mathscr{PR}_\alpha(i)$ and $\mathscr{PRB}_\alpha(i)$ for different Barabási-Albert synthetic networks with $n = 100$ nodes

Summary. We have presented some sharp analytical results for the influence of the personalization vector of the Classic and Biplex PageRank centralities. The relative position of intervals $\mathscr{PR}_\alpha(i)$ and $\mathscr{PRB}_\alpha(i)$ is studied, by proving that they always intersect. Several numerical computations on different families of random networks are included, showing that Biplex PageRank centrality is less controllable than Classic PageRank for dumping factor $\alpha > 0.5$.

References

1. Boccaletti, S., Bianconi, G., Criado, R., del Genio, C.I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks, Physics Reports, 544(1),1–122 (2014)
2. Boldi, P., Santini, M., Vigna, S.: PageRank: Functional Dependencies, ACM Trans. Inf. Syst. 27(4), 19:1–19:23 (2009)
3. Flores, J., García, E., Pedroche F., Romance, M.: Parametric Controllability of the personalized PageRank: classic model vs. biplex approach, Submitted for publication, 1–15 (2019)
4. García, E., Pedroche F., Romance, M.: On the localization of the Personalized PageRank of Complex Networks, Linear Algebra and its Applications 439, 640–652 (2013)
5. Nicosia, V., Criado, R., Romance M., Russo, G., Latora, V.: Controlling centrality in complex networks, Scientific Reports 2, 218 (2012)

6. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bridging order to the Web, Tech.Rep. 66, Stanford University (1998)
7. Pedroche, F., García, E., Romance, M., Criado, R.: Sharp estimates for the personalized Multiplex PageRank, Journal of Computational and Applied Mathematics 330, 1030–1040 (2018)
8. Pedroche, F., Romance, M., Criado, R.: A biplex approach to PageRank centrality: From classic to multiplex networks, Chaos 26, 065301 (2016)

A framework for the construction of generative models for mesoscale structure in multilayer networks

Marya Bazzi*[1,2,3], Lucas G. S. Jeub*[1,4,5], Alex Arenas[6], Sam D. Howison[1], and Mason A. Porter[1,7,8]

[1] Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, United Kingdom
[2] The Alan Turing Institute, London, United Kingdom
[3] Warwick Mathematics Institute, University of Warwick, United Kingdom
[4] Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington
[5] ISI Foundation, Turin, Italy
[6] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Spain
[7] CABDyN Complexity Centre, University of Oxford, United Kingdom
[8] Department of Mathematics, University of California, Los Angeles

One can model many physical, technological, biological, financial, and social systems as networks, which in their simplest form yield graphs [1]. The standard type of network is a single-layer network (also called a "monolayer network"). However, this relatively simple structure cannot capture many of the possible intricacies of connectivity patterns between entities. For example, in temporal networks [2], nodes and/or edges change in time; and in multiplex networks [3], multiple types of interactions can occur between the same pairs of nodes. To better account for the complexity, diversity, and dependencies in real-world interactions, one can represent such connectivity patterns using multilayer networks [3, 4].

Our motivation for considering a single multilayer network instead of several independent single-layer networks is to take into account that connectivity patterns in different layers often depend on each other. Data sets that have multilayer structures are increasingly available (e.g., see Table 2 of [3]). A natural type of multilayer network consists of a sequence of dependent single-layer networks, where layers may correspond to different temporal snapshots, different types of related interactions that occur during a given time interval, and so on. Following existing terminology, we refer to an instance of a node in a layer as a "state node".

Given a (single-layer or multilayer) network representation of a system, it is often useful to apply a coarse-graining technique to investigate features that lie between those at the microscale (e.g., nodes or pairwise interactions between nodes) and those at the macroscale (e.g., total edge weight or degree distribution) [5]. One thereby studies mesoscale features such as community structure [5], core–periphery structure [6], role structure [7], and others. We refer to a set in a network partition that corresponds to some mesoscale structure as a "meso-set" (so community is a type of meso-set).

The ubiquity and diversity of mesoscale structures in empirical networks make it crucial to develop generative models of mesoscale structure that can yield features that one encounters in empirical networks. Broadly speaking, the goal of such generative models is to construct synthetic networks that resemble real-world networks when one appropriately constrains the model parameters using information about the application at hand. Generative models of mesoscale structure can serve a variety of purposes, such as (1) generating benchmark network models for comparing meso-set-detection methods and algorithms [8, 9]; (2) undertaking statistical inference on empirical networks [10, 11]; (3) generating synthetic networks with a desired set of properties [12, 13]; and (4) investigating "detectability limits" for mesoscale structure [9, 14].

One of the main challenges in constructing a realistic generative model (even for single-layer networks) is the breadth of possible empirical features in networks. The available generative models for mesoscale structure in single-layer networks usually focus on replicating a few empirical features at a time (rather than all of them at once): heterogeneous degree distributions and community-size distributions [8, 10], edge-weight distribution [11, 15], spatial embeddedness [16, 18], and so on. Multilayer networks inherit all of the empirical features of single-layer networks, and they also have a key additional one: dependencies between layers. Interlayer dependencies in multilayer networks can be temporal (ordered), multiplex (unordered), or combinations thereof (partially ordered). However, despite this variety, existing generative models for mesoscale structure in multilayer networks allow only a restrictive set of interlayer dependencies (e.g., they assume a temporal structure [9, 16], a simplified multiplex structure with the same planted partitions across all layers [13, 17] or independent groups of layers in which layers in the same group have identical planted partitions [19], etc).

A key feature of multilayer networks is their flexibility, which allows one to incorporate many different types of data in a single structure. In this spirit, we introduce one general framework that enables users of our generative model to construct families of multilayer networks with a range of features of interest in empirical multilayer networks by appropriately constraining the parameter space [20]. To achieve this, we choose to first partition the set of state nodes of a multilayer network and we then allocate edges, given a multilayer partition. Specifically, we focus on modeling dependency at the level of partitions (as was done in [9]), rather than with respect to edges; and we treat the process of generating a multilayer partition separately from that of generating edges for a given multilayer partition. This modular

The 8th International Conference on Complex Networks and Their Applications, 10 - 12 Dec. 2019, Lisbon, Portugal

approach, which enables the use of all existing network models with a planted partition, yields random structures that can capture a wide variety of interlayer-dependency structures (e.g., temporal and/or multiplex networks, appearance and/or disappearance of entities, uniform or nonuniform dependencies between state nodes from different layers, and so on). For a specified interlayer-dependency structure, one can then use any network model with a planted partition to generate a wide variety of network features, including weighted edges, directed edges, and spatially-embedded layers.

The flexibility of our model to generate multilayer networks with a specified dependency structure between different layers makes it possible to (1) gain insight into whether, when, and how to build interlayer dependencies into methods for studying different types of multilayer networks; and (2) generate tunable benchmarks to allow a principled comparison for community-detection (and, more generally, meso-set-detection) tools for multilayer networks.

1 Results

We introduce a general and customisable generative model for mesoscale structures in multilayer networks [20]. The complexity of dependencies between layers can make it difficult to explicitly specify a joint probability distribution for meso-set assignments, especially for unordered or partially ordered multilayer networks. To address this issue, we define a conditional probability model on a state node's meso-set assignment, given the assignments of all other state nodes. Specifying conditional models (which capture different dependency features separately) rather than joint models (which try to capture many dependency features at once) is convenient for numerous situations. We parametrise the conditional partition model with two key parameters: (1) layer-specific null distributions and (2) an interlayer dependency tensor. The former allows the incorporation of certain desirable features for any choice of interlayer dependency (e.g., variation in the expected number and sizes of meso-sets across layers) and the latter allows the explicit parametrisation of dependencies between different layers. Using the conditional model, we define an iterative copying process on the meso-set assignments of state nodes to generate multilayer partitions with dependencies between induced partitions in different layers.

Consider a node j in layer β and let V_M be the set of state nodes in a multilayer network. We denote the user-specified interlayer-dependency tensor by $\mathbf{P}$, where $P_{i,\alpha}^{j,\beta}$ is the probability that state node (j,β) copies its meso-set assignment from state node (i,α), for any two state nodes $(i,\alpha),(j,\beta) \in V_M$. The interlayer-dependency tensor induces the interlayer-dependency network, whose edges are all interlayer, directed, and pointing in the direction of information flow between layers. The probability that state node (j,β) copies its meso-set assignment from an arbitrary state node when state node (j,β)'s meso-set assignment is updated is

$$\hat{p}_{j,\beta} = \sum_{(i,\alpha)\in V_M} P_{i,\alpha}^{j,\beta} \,,$$

where we require that $\hat{p}_{j,\beta} \leq 1$ for all state nodes $(j,\beta) \in V_M$. Suppose that we are updating the meso-set assignment of state node (j,β) at step τ of the copying process and that the current multilayer partition is $\mathbf{S}(\tau)$. With probability $\hat{p}_{j,\beta}$, a state node (j,β) copies its meso-set assignment from one of its in-neighbors in the interlayer-dependency network; and with probability $1-\hat{p}_{j,\beta}$, it obtains its meso-set assignment from the null distribution $\mathbb{P}_0^{\beta}$. This yields the following update equation at step τ of our copying process:

$$\begin{aligned} \mathbb{P}[S_{j,\beta}(\tau+1) = s | \mathbf{S}(\tau)] \\ = \sum_{(i,\alpha)\in V_M} P_{i,\alpha}^{j,\beta}\, \delta(S_{i,\alpha}(\tau), s) \\ + \left(1-\hat{p}_{j,\beta}\right) \mathbb{P}_0^{\beta}[S_{j,\beta} = s] \,. \end{aligned} \tag{1}$$

The update equation (1) is at the heart of our generative model. It is clear from (1) that the set of null distributions is responsible for the specification of meso-set assignments in the absence of interlayer dependencies (i.e., if $P_{i,\alpha}^{j,\beta} = 0$ for all (i,α), (j,β)). In general, $\hat{p}_{j,\beta}$ determines the relative importance of interlayer dependencies and the null distribution on the meso-set assignment of state node (j,β). Specifically, when $\hat{p}_{j,\beta} = 0$, the meso-set assignment of (j,β) depends only on the null distribution; and when $\hat{p}_{j,\beta} = 1$, the meso-set assignment of (j,β) depends only on the meso-set assignments of its in-neighbors in the interlayer-dependency network.

When updating the meso-set assignments of state nodes, we respect the order the layers (e.g., temporal ordering). For a fully ordered multilayer network (e.g., temporal), our update process reduces to sequentially sampling an induced partition for each layer based on the induced partitions of previous layers. For an unordered multilayer network (e.g., multiplex), our update process defines a Markov chain on the space of multilayer partitions. Our sampling strategy reduces to (pseudo-)Gibbs sampling [21, 22], an approach in which one samples partitions from a stationary distribution of this Markov chain. For a partially ordered multilayer network (e.g., multiplex network that changes over time), our update

process combines these two sampling strategies. We discuss the parameters and properties of our generative model, and we illustrate examples of its use with benchmark models for community-detection methods and algorithms in multilayer networks [20].

The three most important features of our model are the following: (1) it includes an explicitly parametrizable tensor that controls interlayer-dependency structure; (2) it can generate an extremely general, diverse set of multilayer networks (including, e.g., temporal and/or multiplex); and (3) it is modular, as the process of generating a partition is separate from the process of generating edges, enabling a user to first generate a partition and then use any planted-partition network model. We provide publicly available code (https://github.com/MultilayerGM) that users can modify to readily incorporate different types of null distributions, interlayer-dependency structures, and planted-partition network models.

Summary. Multilayer networks allow one to represent diverse and coupled connectivity patterns — e.g., time-dependence, multiple subsystems, or both — that arise in many applications and which are difficult or awkward to incorporate into standard network representations. In the study of multilayer networks, it is important to investigate mesoscale (i.e., intermediate-scale) structures, such as dense sets of nodes known as communities, to discover network features that are not apparent at the microscale or the macroscale. We introduce a generative model for mesoscale structure in multilayer networks. Our model is very general, with the ability to produce many features of empirical multilayer networks, and it explicitly incorporates a user-specified dependency structure between layers. Our results provide a standardized set of null models, together with an associated set of principles from which they are derived, for studies of mesoscale structures in multilayer networks. We discuss the parameters and properties of our generative model, and we illustrate examples of its use with benchmark models for community-detection methods and algorithms in multilayer networks.

References

1. M. Newman: Networks. Oxford University Press, 2018.
2. P. Holme: Modern temporal network theory: A colloquium. Eur. Phys. J. B., 2015.
3. M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno, and M. A. Porter: Multilayer networks. J. Complex Netw., 2014.
4. G. Bianconi: Multilayer networks: structure and function. Oxford University Press, 2018.
5. M. A. Porter, J.-P. Onnela and P. J. Mucha: Multilayer networks: structure and function. Notices Amer. Math. Soc., 2009.
6. P. Rombach, M. A. Porter, J. H. Fowler and P. J. Mucha: Core–periphery structure in networks (Revisited). SIAM Rev., 2017.
7. R. A. Rossi and N. K. Ahmed: Role Discovery in Networks. IEEE Trans. Knowl. Data Eng., 2015.
8. A. Lancichinetti, S. Fortunato and F. Radicchi: Benchmark graphs for testing community detection algorithms. Phys. Rev. E, 2008.
9. A. Ghasemian, P. Zhang, A. Clauset, C. Moore and L. Peel: Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks. Phys. Rev. X, 2016.
10. B. Karrer and M. E. J. Newman: Stochastic blockmodels and community structure in networks. Phys. Rev. E, 2011.
11. T. P. Peixoto: Nonparametric weighted stochastic block models, Phys. Rev. E, 2018.
12. L. G. S. Jeub, O. Sporns, and S. Fortunato: Multiresolution Consensus Clustering in Networks, Sci. Rep., 2018.
13. C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore: Community detection, link prediction, and layer interdependence in multilayer networks, Phys. Rev. E, 2017.
14. D. Taylor, R. S. Caceres, and P. J. Mucha: Super-resolution community detection for layer-aggregated multilayer networks, Phys. Rev. X, 2017.
15. C. Aicher, A. Z. Jacobs, and A. Clauset: Learning latent block structure in weighted networks, J. Complex Netw., 2014.
16. M. Sarzynska, E. A. Leicht, G. Chowell, and M. A. Porter: Null Models for Community Detection in Spatially-Embedded, Temporal Networks, J. Complex Netw., 2015.
17. T. P. Peixoto: Inferring the mesoscale structure of layered, edge-valued, and time-varying networks, Phys. Rev. E, 2015.
18. M. E. J. Newman and T. P. Peixoto: Generalized Communities in Networks, Phys. Rev. Lett., 2015.
19. N. Stanley, S. Shai, D. Taylor, and P. J. Mucha: Clustering network layers with the strata multilayer stochastic block model, IEEE Trans. Network Sci. Eng., 2016.
20. M. Bazzi, L. G. S. Jeub, A. Arenas, S. D. Howison, and M. A. Porter: A framework for the construction of generative models for mesoscale structure in multilayer networks, arXiv 1608.06196, 2019.
21. A. E. Gelfand: Gibbs Sampling, J. Am. Stat. Assoc., 2000.
22. D. Heckerman, D. M. Chickering, C. Meek, R. Roundthwaite, C. Kadie: Dependency Networks for Inference, Collaborative Filtering, and Data Visualization, J. Mach. Learn. Res., 2000.

Autoencoders and Graph Convolutional Networks for Multilayer Network Embedding

Diego Perna[1], Roberto Interdonato[2], and Andrea Tagarelli[1]

[1] DIMES, University of Calabria, Italy. [2] CIRAD, UMR Tetis, Montpellier, France.
d.perna@dimes.unical.it, roberto.interdonato@cirad.fr,
andrea.tagarelli@unical.it

1 Introduction

Network embedding methods [7–10] have attracted increasing attention as key-enabling tool to successfully address emerging challenges in large real-world networks (such as high computational complexity, low parallelizability, and inapplicability of machine learning methods) by learning a low-dimensional representation of one or more components of a graph network. In particular, deep embedding methods, such as those based on convolutional neural networks, showed their effectiveness on a broad range of problems in different fields (e.g., speech and image recognition), and they promise to achieve unprecedented opportunities also in network science.

However, such methods have traditionally focused on structured data (e.g., grids), while there is an inherent difficulty in defining basic operations, such as convolution [3], in graph networks; in fact, defining the convolution operation on grid-structured data is straightforward (e.g., each pixel in an image can be seen as an element, and the size of its neighborhood is determined by the size of the kernel), whereas in the case of graphs, nodes are unordered and the size of their neighborhood can vary largely. One way of performing the convolution operation on graph data is to aggregate the values of each node's features along with its neighbors' features. This is the basic approach adopted by the Graph Convolutional Network (GCN) method proposed in [4]. Alternatively, the GraphEncoder method in [6] exploits stacked sparse autoencoders, which have shown to be very similar to spectral clustering in theory yet much more efficient and flexible in practice. It should be noted that *the above methods work on simple networks only*.

In this work, we take inspiration from GraphEncoder, but differently from it we leverage graph convolutional networks and their use in an autoencoder framework [5]. To the best of our knowledge, we are the first to propose an autoencoder-based GCN architecture for learning a compressed representation (i.e., node embeddings) of a multilayer network. Our closely related work is the one recently proposed in [1], which learns two embeddings for each actor: the one obtained by aggregating information from the different layers of the multiplex, and the other one for each node of the multiplex. The two embeddings are then linked together by projection matrices that constrain the generation of layer-specific representations conditioned to the across-layer ones. Note that, however, this linkage relies on a hyper-parameter which adds a further degree of freedom in learning an embedding; also, unlike our proposal, it misses the advantages coming from autoencoders and, particularly, from *variational autoencoders*, which act as generative models to learn the parameters of a probability distribution representing the network data.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Our proposed mGCNAE and mGCNVAE methods

Given a set $\mathcal{V}$ of N *entities* (e.g., users) and a set $\mathcal{L} = \{L_1, \cdots, L_\ell\}$ of *layers* (e.g., user relational contexts), with $\ell \geq 2$, we denote a multilayer network with $G_\mathcal{L} = (V_\mathcal{L}, E_\mathcal{L}, \mathcal{V}, \mathcal{L})$, where $V_\mathcal{L} \subseteq \mathcal{V} \times \mathcal{L}$ is the set of entity-layer pairings or *nodes* (to denote, e.g., each user is present in which layers), and $E_\mathcal{L} \subseteq V_\mathcal{L} \times V_\mathcal{L}$ is the set of undirected edges between nodes within and across layers.

We represent a multilayer network by a set of adjacency matrices $\mathcal{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_\ell\}$, with $\mathbf{A}_l \in \mathbb{R}^{n_l \times n_l}$ ($l = 1..\ell$), where $n_l = |V_l|$. Entities may be associated with *features* stored in layer-specific matrices $\mathcal{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_\ell\}$, with $\mathbf{X}_l \in \mathbb{R}^{n_l \times f_l}$ and f_l the number of node features in the l-th layer. In case no side-information is available for $G_\mathcal{L}$, each layer-specific feature matrix is assumed to be the identity matrix $\mathbf{I}_l \in \mathbb{R}^{n_l \times n_l}$. Note that, since we need to account also for inter-layer edges, the aggregation of features should be computed not only with the ones of each node's neighbors, but also with the features of the different nodes coupled to the same entity over the layers of the multilayer network.

To enable effective convolution, each layer matrix is symmetrically normalized after adding self-loops for all nodes; formally, for the l-th layer, we have $\hat{\mathbf{A}}_l = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}}_l \tilde{\mathbf{D}}^{-\frac{1}{2}}$, with $\tilde{\mathbf{A}}_l = \mathbf{A}_l + \mathbf{I}_l$, $\tilde{\mathbf{D}}_l = \mathbf{D}_l + \mathbf{I}_l$ and $\mathbf{D}_l$ the degree matrix. As shown in [4], this helps shrink the underlying graph spectrum, and as a consequence, nearby nodes will tend to share similar representations. Note that the i-th entry in $\mathbf{D}_l$ stores the degree of entity v_i internal to the l-th layer, plus the number of inter-layer edges that are incident to v_i.

Let us utilize subscript l to denote the l-th layer of the input multilayer network, and the superscript (k) to denote the k-th convolutional layer of the neural system. For the k-th convolutional layer, we denote the input representations of all nodes with the matrix $\mathbf{H}^{(k-1)}$ and the output representation with $\mathbf{H}^{(k)}$, which is usually smaller than $\mathbf{H}^{(k-1)}$. Note that the initial representation is the input feature matrix, i.e., $\mathbf{H}^{(0)} \equiv \mathbf{X}$, which hence feds the first convolutional layer.

Our first proposed approach, dubbed Multilayer Graph Convolutional Network Autoencoder (mGCNAE), requires at first the following embedding rule for every l-th layer of the multilayer network:

$$\mathbf{H}_l^{(1)} = ReLU(\hat{\mathbf{A}}_l drop(\mathbf{X}_l)\mathbf{W}_l^{(1)}) \tag{1}$$

where $\mathbf{W}$ denotes a *weight* matrix in the convolutional layer, the nonlinear activation function $ReLU(\cdot)$ is applied pointwise, and $drop(\cdot)$ is the dropout function typically introduced to help reduce overfitting. Matrix $\mathbf{H}_l^{(1)}$ has dimension $\mathbb{R}^{n_l \times d_1}$ (by default, $d_1 = 32$ [5]), and the layers $\mathcal{L}$ are treated independently.

From the second layer of mGCNAE, an analogous operation is performed, however we devise two different modes and, consequently, two different representation update rules, for every $k = 2..K$:

- handle the different layers of the multilayer graph independently:

$$\mathbf{H}_l^k = \hat{\mathbf{A}}_l drop(\mathbf{H}_l^{(k-1)})\mathbf{W}_l^{(k)} \tag{2}$$

- share a unique weight matrix $\mathbf{W}^{(k)}$ through the different layers of the multilayer graph:

$$\mathbf{H}_l^k = \hat{\mathbf{A}}_l drop(\mathbf{H}_l^{(k-1)})\mathbf{W}^{(k)} \tag{3}$$

Finally, the output of the K-th hidden layer corresponds to the new learned node representation (i.e., the embedding), $\mathbf{Z} = \mathbf{H}^{(K)}$. Once obtained the embedding, the decoding phase consists in the inner product $\mathbf{Z}_l \mathbf{Z}_l^T$ to achieve the reconstructed adjacency matrix $\bar{\mathbf{A}}_l$ for the l-th layer of the multilayer graph.

Our second method, dubbed mGCNVAE, exploits the variational autoencoder paradigm, so that it differs from mGCNAE in that the second hidden layer consists of two components which constitute the core of the inference model:

$$\mathbf{Z}_l^{(\mu)} = ReLU(\hat{\mathbf{A}}_l drop(\mathbf{H}_l^{(1)})\mathbf{W}_l^{(\mu)}) \text{ and } \mathbf{Z}_l^{(\sigma)} = ReLU(\hat{\mathbf{A}}_l drop(\mathbf{H}_l^{(1)})\mathbf{W}_l^{(\sigma)}) \quad (4)$$

with μ (resp. σ) mean (resp. standard deviation) of the latent representation. The encoding step corresponds to a function q to learn:

$$q(\mathbf{Z}_l|\mathbf{X}_l,\mathbf{A}_l) = \prod_{i=1}^{N} q(\mathbf{z}_{l,i}|\mathbf{X}_l,\mathbf{A}_l), \text{ with } q(\mathbf{z}_{l,i}|\mathbf{X}_l,\mathbf{A}_l) = \mathcal{N}(\mathbf{z}_{l,i}|\mathbf{z}_{l,i}^{(\mu)}, diag(\mathbf{z}_{l,i}^{(\sigma^2)})) \quad (5)$$

while the generative model (i.e., the decoder) is given by function p:

$$p(\mathbf{A}_l|\mathbf{Z}_l) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(A_{l,ij}|\mathbf{z}_{l,i}\mathbf{z}_{l,j}), \text{ with } p(A_{l,ij}=1|\mathbf{z}_{l,i}\mathbf{z}_{l,j}) = \sigma(\mathbf{z}_{l,i}^T\mathbf{z}_{l,j}) \quad (6)$$

and the loss function is: $\sum_{l=1}^{\ell} \mathbb{E}_{q(\mathbf{Z}_l|\mathbf{X}_l,\mathbf{A}_l)}[\log p(\mathbf{A}_l|\mathbf{Z}_l)] - KL[q(\mathbf{Z}_l|\mathbf{X}_l,\mathbf{A}_l)||p(\mathbf{Z}_l)]$.

3 Preliminary experimental results

We implemented our proposed methods and tested them, for a *link prediction* task, on a number of multilayer networks, including Cora and Citeseer [16, 17], StarWars [18], London transportation [15], EUair [14], FAO [13], ArXiv [12], and Pierre Auger [12]. We randomly selected and removed 5% (resp. 10%) of the edges from the input multilayer graph to be used for validation (resp. testing). Next, we integrated validation and testing edge sets with unconnected pairs of nodes in a balanced fashion. Once obtained the embeddings, we assessed the similarity between pairs of nodes to predict whether the edge between the two nodes existed. For the sake of simplicity, we assumed n_l and f_l to be the same for all layers, and used two convolutional hidden layers ($K = 2$). Moreover, we devised 4 settings of the two methods, depending on whether only intra-layer edges or also inter-layer edges were used by the neural system, and whether the across-layer shared weight matrix was used or not.

Preliminary results show that, on average over all datasets, mGCNAE and mGCNVAE obtained good performance in terms of AUC (0.73), precision (0.78) and accuracy (0.65), with peaks on the Pierre Auger dataset by m-GCNVAE in terms of AUC (0.98) and precision (0.98), and by m-GCNAE in terms of accuracy (0.93), for the only-intra-layer-edges variants; also on the two citation networks, any variant of both methods performed very well. By contrast, the variants using both intra- and inter-layer edges appeared to perform worse, except on EUair; the latter network, on the other hand, represents the more difficult testbed, due to their higher number of layers and their extremely varying structure [14]. Besides that, in general, using both intra- and inter-layer edges may be more affected by the lack of relevant (side-)information associated with inter-layer edges.

References

1. Ma, Y., Wang, S., Aggarwal, C. C., Yin, D., & Tang, J. (2019). Multi-dimensional Graph Convolutional Networks. In Proc. of the SIAM Int. Conf. on Data Mining (pp. 657-665).
2. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.
3. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4), 18-42.
4. Thomas N. Kipf, Max Welling (2017), Semi-Supervised Classification with Graph Convolutional Networks. In Proc. of Int. Conf. on Learning Representations (ICLR).
5. Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders. In Proc. NeurIPS Workshop on Bayesian Deep Learning.
6. Tian, F., Gao, B., Cui, Q., Chen, E., & Liu, T. Y. (2014). Learning deep representations for graph clustering. In Proc. AAAI Conf. on Artificial Intelligence.
7. Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 151, 78-94.
8. Cui, P., Wang, X., Pei, J., & Zhu, W. (2018). A survey on network embedding. IEEE Trans. on Knowledge and Data Engineering, 31(5), 833-852.
9. Cai, H., Zheng, V. W., & Chang, K. C. C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Trans. on Knowledge and Data Engineering, 30(9), 1616-1637.
10. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2019). A comprehensive survey on graph neural networks. arXiv preprint arXiv:1901.00596.
11. Vickers, M., & Chan, S. (1981). Representing classroom social structure. Victoria Institute of Secondary Education, Melbourne.
12. De Domenico, M., Lancichinetti, A., Arenas, A., & Rosvall, M. (2015). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. Physical Review X, 5(1), 011027.
13. De Domenico, M., Nicosia, V., Arenas, A., & Latora, V. (2015). Structural reducibility of multilayer networks. Nature Communications, 6, 6864.
14. Cardillo, A., Gmez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Del Pozo, F., & Boccaletti, S. (2013). Emergence of network features from multiplexity. Scientific reports, 3, 1344.
15. De Domenico, M., Sol-Ribalta, A., Gmez, S., & Arenas, A. (2014). Navigability of interconnected networks under random failures. Proceedings of the National Academy of Sciences, 111(23), 8351-8356.
16. Lu, Q., & Getoor, L. (2003). Link-based classification. In Proc. of Int. Conf. on Machine Learning (ICML) (pp. 496-503).
17. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. AI magazine, 29(3), 93-93.
18. Gabasova, E. (2015). The Star Wars social network. Evelina Gabasova's blog. https://github.com/evelinag/StarWars-socialnetwork/tree/master/networks.

A resilience trade-off for inter-layer connectivity in multimodal transportation networks

Camill Harter[1], Otto Koppius[1], Rob Zuidwijk[1]

Rotterdam School of Management, Erasmus University,
Rotterdam, The Netherlands,
harter@rsm.nl

1 Motivation

Interdependent complex systems have been shown to be highly vulnerable to attack as failures cascade across layers leading to a rapid disintegration of the network[1, 2]. The risk of cascades of failure is created by inter-layer connections, which represent the dependencies between the network layers. Consequently, most studies on resilience of interdependent networks consider inter-layer connections a necessary evil needed to sustain a systems functionality, but detrimental to its resilience[3, 4].

While this makes sense for certain types of networks such as the power-communication interdependent system, which played a crucial role in the famous 2003 blackout in Italy, it does not hold for all types of interdependent networks. In certain network contexts, inter-layer connectivity does not only present a risk, but also a chance for resilience[5].

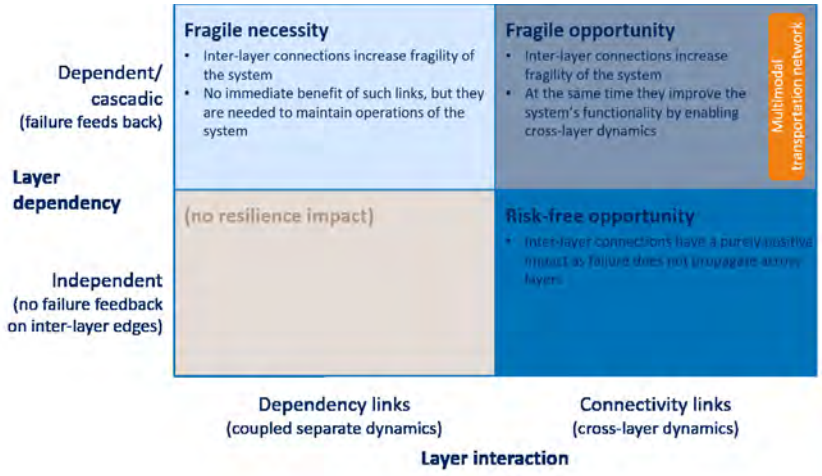


Fig. 1. Illustration of resilience impact of inter-layer edges in multiplex networks depending on the functional coupling (layer interaction and dependency)

Inter-layer connectivity presents a chance for network resilience if inter-layer edges have a function that allows for collaboration and shared dynamics between the layers,

and do not only describe a dependency relation. For instance, strategic interconnectivity between isolated power grids can suppress cascades of failure as excessive load can better be distributed across different grids. However, too much interconnectivity increases the risk of larger cascades infecting multiple power grids[5]. Furthermore, in multimodal transport networks, mode transshipment links increase flexibility to use different transport modes in response to disruption, thereby mitigating its impact. At the same time, intensifying inter-layer coupling increases the risk of cascading failure.

This suggests that inter-layer connections can have a very different impact on the resilience of a multiplex network depending on the type of functional interaction and dependency between layers. Figure 1 shows a framework describing the different categories of functional coupling and the resilience impact of inter-layer edges for each category.

The focus of this study lies on the 'Fragile opportunity' category. For networks in this category, inter- layer edges create benefits and make them more fragile at the same time. This suggests that there is a trade-off in the layer coupling structure between topological stability (risk of cascadic failure) and operational functionality. This trade-off comes into play as density and location of inter-layer edges vary. Thus, in order to understand it, the structural coupling of layers needs to be analyzed. Therefore, different synthetic multiplex networks varying in the structure of their layers are analysed and their resilience against disruption under different coupling structures (intensity of coupling and coupling pattern) is assessed. Moreover, we study the resilience of the European hinterland transport network for container shipping, a multiplex transportation network with layers formed by transport modes rail and barge.

The results contribute to the understanding how inter-layer connectivity can influence resilience in different types of multiplex networks and how inter-layer connections should be chosen to foster resilience and mitigate the impact of disruption. This is important for decision makers in hinterland shipping as transport modes are becoming more strongly coupled towards an integrated intermodal system. For our analysis we make use of a unique dataset containing all rail and barge services scheduled in the European hinterland. We create a multiplex network with nodes representing cities that have at least one container terminal and edges representing transport connections. Layers are formed by the two alternative transport modes.

2 Initial results

As a first step, the impact of coupling intensity, i.e. the share q of overlapping nodes that is linked by an inter-layer edge, on resilience is assessed. Therefore, a multiplex network with layers formed by two Erdos-Renyi networks with 100 nodes and $p = 0.053$ is studied. The two layers are partially overlapping at 50% of their nodes ($Q = 0.5$). Initial failure of nodes is done randomly and cascadic failure propagation is modeled as in [1]. Resilience is measured by the change in network efficiency [6], which is a suitable measure particularly for transportation networks.

Figure 2 shows the results for three different initial attack sizes (share of nodes $p \in \{0, 0.2, 1\}$). If no attack takes place ($p = 0$), inter-layer edges have a purely positive effect as they connect the layers and distances become shorter. The resilience trade-

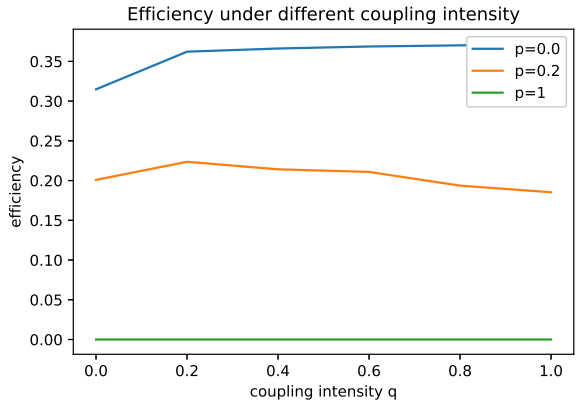


Fig. 2. Efficiency of barge-rail coupled network for hinterland transport in Europe depending on inter-layer coupling intensity q and initial attack sizes $p \in \{0, 0.2, 1\}$

off becomes visible at intermediate attack sizes ($p = 0.2$). At low coupling intensities, each added inter-layer edge improves network efficiency as routing becomes easier. However, the marginal added value of an additional edge decreases and at a certain point, efficiency declines as the network becomes more and more fragile. This point marks the optimal coupling intensity as resilience reaches its maximum under the given network settings. Results can vary strongly depending on the choice of network type, attack strategy, and coupling strategies.

References

1. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. Nature, 464(7291), 1025.
2. Gao, J., Barzel, B., Barabsi, A. L. (2016). Universal resilience patterns in complex networks. Nature, 530(7590), 307.
3. Parshani, R., Buldyrev, S. V., Havlin, S. (2010). Interdependent networks: Reducing the coupling strength leads to a change from a first to second order percolation transition. Physical review letters, 105(4), 048701.
4. Schneider, C. M., Yazdani, N., Arajo, N. A., Havlin, S., Herrmann, H. J. (2013). Towards designing robust coupled networks. Scientific reports, 3, 1969.
5. Brummitt, C. D., DSouza, R. M., Leicht, E. A. (2012). Suppressing cascades of load in interdependent networks. Proceedings of the National Academy of Sciences, 109(12), E680-E689.
6. Latora, V., Marchiori, M. (2001). Efficient behavior of small-world networks. Physical review letters, 87(19), 198701.

Part X

Network Analysis and Measure

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

A Proposal for the E-I Index for Non-disjoint Groups

Ricardo Lopes de Andrade[1] and Leandro Chaves Rêgo[1,2]

[1] Production Engineering Graduate Program
Universidade Federal de Pernambuco
Recife/PE, 50740-550, Brazil,
ricardolopesa@gmail.com,
[2] Statistics and Applied Math Department
Universidade Federal do Ceará
Fortaleza/CE, 60440-900, Brazil,
leandro@dema.ufc.br

1 Introduction

The studies on homophilia in social networks seek to quantify the propensity of individuals to interact with similar actors ([1], [2], [3]). In these studies, the E-I index proposed by Krackhardt and Stern (1988) [4], is used as a measure for homophilia. The E-I index is a simple measure obtained from the difference between the number of external links (links between nodes belonging to different groups - EL) and the number of internal links (links between nodes belonging to the same group - IL), divided by the total number of connections for normalization.

$$E-I\ Index = \frac{EL-IL}{EL+IL} \tag{1}$$

The E-I index ranges from -1 (all bonds are internal) to +1 (all bonds are external). The index can be calculated for the entire network, for each group or for each individual actor. In a weighted network, EL is the sum of the edge's weights that connect different cells of the partition and IL is the sum of the edge's weights that connect actors of the same cell of the partition.

The nodes of the network are assigned to groups, for example: age-based [1]; based on ethnicity [3]; based on gender, religion, politics [5]; among others. Grouping involves partitioning the set of nodes into exhaustive and mutually exclusive subsets. Publications that use the E-I index as a measure of homophilia or segregation are concentrated in disjoint or mutually exclusive groups, that is, each node or actor in the network has only one bond of a particular attribute. Situations where network actors are present in more than one group, such as non-disjoint groups, are not commonly explored. However, distinctive disjoint groups rarely exist at large scales in many empirical networks [6], what is observed in many analyzes, in fact, is a set of overlapping groups rather than partitions [7]. Some attributes such as: areas of knowledge in networks of researchers; economic blocks in commercial networks; communities in networks such as Facebook, Twitter, among others; and other attributes linked to behaviors, tastes and attitudes generate non-disjoint groups. One of the barriers encountered in the analysis of non-disjoint groups is the absence of a measure, since the E-I index is defined for disjoint groups.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

In this context, the objective of this work is to develop a measure that quantifies the relational structure within and between groups that encompass not only the analysis of disjoint groups but also non-disjoint groups. Allowing the expansion of the analysis of social networks, for several types of attributes, helps identifying which actors have more similarities or differences, generating previously unexploited knowledge. Specifically, we generalize the E-I index developed by [4] to deal with non-disjoint groups.

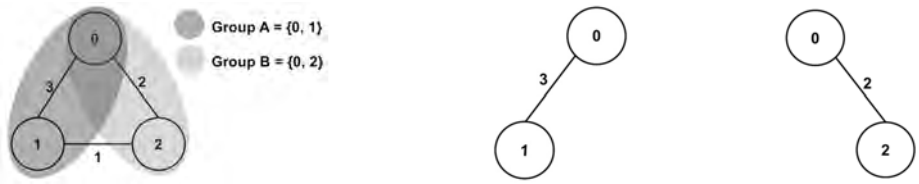
2 Results

In order to explore cases of non-disjoint groups, we have developed a new method to obtain the E-I index, which is a generalization of the current method. Figure 1 is used to illustrate the new method. Figure 1a has three nodes and two generic attribute groups, nodes 0 and 1 have attribute A and nodes 0 and 2 have attribute B. Therefore, both attribute groups have one node in common. In Figures 1b and 1c, nodes are connected if they belong to the same group. In this context, the E-I Index is defined as follows:

- For a set of nodes: EL is the number of nodes' edges in the first graph that are not present in any of the group graphs and IL is the number of nodes' edges in the first graph that are present in at least one of the group graphs.
- For an attribute group: EL is the number of node edges of a given group in the first graph that are not present in the given group graph and IL is the number of node edges of a given group in the first graph that are present in the given group graph.

Table 1: E-I Index non-disjoint group example

	Unweighted			Weighted		
	EL	IL	E-I Index	EL	IL	E-I Index
set {0}	0	2	-1.00	0	5	-1.00
set {1}	1	1	0.00	1	3	-0.50
set {2}	1	1	0.00	1	2	-0.33
set {0,1,2}	1	2	-0.33	1	5	-0.66
Group A	2	1	0.33	3	3	0.00
Group B	2	1	0.33	4	2	0.33



(a) Graph G (b) Graph A (c) Graph B

Fig. 1: Social network with non-disjoint groups of nodes.

Table 1 displays the results for the graph shown in Figure 1a. It is easy to verify that the proposed metric is a generalization of the E-I index proposed in [4] in the sense that if groups are disjoint, then it coincides with (1).

In future work, we will implement the proposed model in two real networks:

(i) Co-authorship PQ: The PQ network is a co-authorship network among researchers in the area of Industrial Engineering of Brazil, has 92 nodes in the giant component and 131 edges. The network is undirected and the edges represent the publications in co-authorship [8].The sets of overlapping groups are the industrial engineering areas of knowledge.
(ii) Trade of American Countries: The network of commerce between the American countries is formed 30 countries and 356 edges. This network is a subnetwork of the network of international trade, developed by [9], which includes 178 countries that form a unique main component with 10,419 edges. The network is undirected and the edges represent the commercial transactions between countries. The sets of overlapping groups are the blocks or trade agreements that the American countries are inserted into.

As these examples suggest, in applications, the groups do have an empirical meaning and the E-I index is just a way to classify sets of nodes or attribute groups according to the proportion of their internal/external links.

Acknowledgments: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. The first and second authors would like to acknowledge the financial support of the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) and of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), respectively.

References

1. Burt RS. Measuring age as a structural concept. Soc Networks 1991 doi:10.1016/0378-8733(91)90011-H
2. McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in Social Networks. Annu Rev Sociol (2001) doi:10.1146/annurev.soc.27.1.415
3. Alvarez Benjumea, Amalia. "Homophily and Ethnic Background in the Classroom." (2015).
4. Krackhardt D, Stern RN. Informal Networks and Organizational Crises: An Experimental Simulation. Soc Psychol Q (1988) doi:10.2307/2786835
5. Saeidibonab, Sepehr. "Homophily and Friendship Dynamics: An analysis of friendship formation with respect to homophily principle and distinctiveness theory." (2017)
6. Leskovec, Jure, et al. Statistical properties of community structure in large social and information networks. Proceedings of the 17th international conference on World Wide Web. ACM, (2008)
7. Everett, Martin G., and Stephen P. Borgatti. "Categorical attribute based centrality: EI and GF centrality." Social Networks 34.4: 562-569 (2012) doi.org/10.1016/j.socnet.2012.06.002
8. Andrade RL de, Rêgo LC. Exploring the co-authorship network among CNPqs productivity fellows in the area of industrial engineering. Pesqui Operacional (2017) doi:10.1590/0101-7438.2017.037.02.0277
9. Andrade RL de, Rêgo LC. The use of nodes attributes in social network analysis with an application to an international trade network. Phys A Stat Mech Its Appl (2018) doi:10.1016/j.physa.2017.08.126

Disentangling Public Transit Ridership into a Spatiotemporal Geography

Mikhail Sirenko[1], Scott Cunningham[1,2], Nuno A. M. Araújo[3], and Trivik Verma[1,4]

[1] Faculty of Technology, Policy and Management, Delft University of Technology
t.verma@tudelft.nl
[2] Faculty of Humanities and Social Science, University of Strathclyde, Scotland
[3] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Portugal
[4] Environmental Systems Science, ETH Zurich

1 Introduction

The spaces and networks which comprise a city evolve in a complex and interdependent manner. Urban spaces are occupied by an increasing diversity of citizens, all with varying needs and requirements. Transport networks are increasing in form and variety and play a central role in providing enhanced access, among and across neighbourhoods, to residential places, working institutions and local amenities like shopping malls, restaurants and hospitals. Improved accessibility through transportation networks also threatens to create larger, more sprawling urban areas [1].

Discerning the urban structure in the context of public transit utilisation is significant for urban planning [2] and sometimes central in controlling the urban sprawl [3] through Transit-oriented Development (TOD). In urban planning research, urban structures are generally characterised using origin-destination (OD) matrices that represent individual and cumulative mobility flows in a city [4]. A standard sociological method involves survey-based direct observation of urban populations [5] and their transport patterns such as walking and vehicle ownership [6] or more novel analyses of mobile telephony networks [7]. However, such indicators are merely representative of formal or designated use of the "diurnal" and "bimodal" city that are leveraged for standard planning applications. Moreover, this framework of estimating demand and developing supply is heavily model-based, relying on several parameters, and difficult to analyse and compare through time. Given this, new methods are needed to assess the multiple types and spatial flows of passengers on urban transportation networks.

Digital services like Automatic Fare Collection (AFC) have been introduced into transit networks worldwide and enable an unprecedented amount of anonymous transit ridership data. We aim to illustrate a simple and powerful method with an example where entry-only ridership data of $\approx 4.5M$ daily traces from the Greater London region can be transformed into spatiotemporal geography of the city. Each station decomposes into a mix of *six* distinct ridership classes across time, while simultaneously being classified in spaces of central, polycentric and concentric development. Our method can be applied to any region in

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the world where entry-only ridership data is available and could be useful for data-driven planning.

2 Results

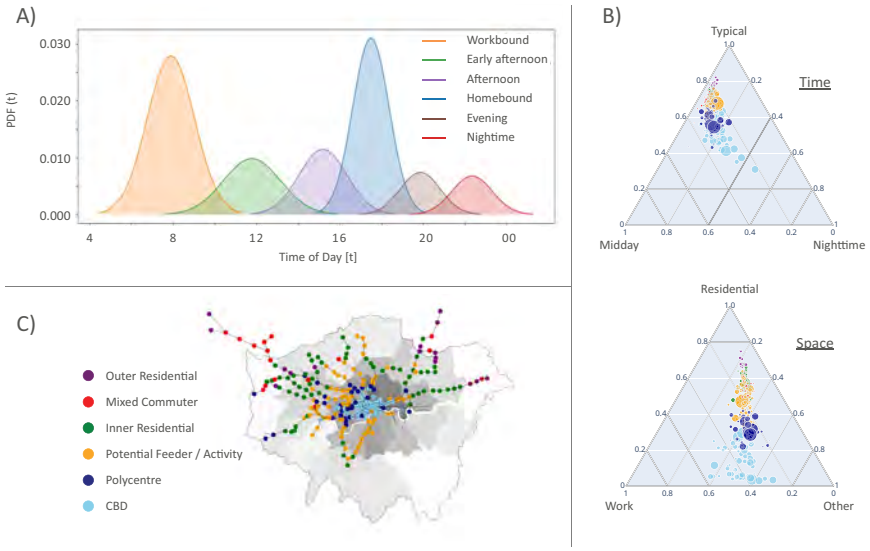


Fig. 1. Disentangling Public Transit Ridership into Time and Space. A). Probability Density Function of the individual ridership data of Greater London decomposed into its characteristic mixtures across a day using GMM, with $K = 6$. B). Ternary plots showing composite behaviour of stations with the mixtures either characterising stations in time or space. Typical commuters are work or homebound. Midday commutes relate to afternoon traffic. Nighttime commutes are following standard work traffic. C). A population map describing the spatial classification of stations into six types, each type composed of a percentage mix of mixtures as shown in part (A).

Public Transit Ridership. We use the London Underground Passenger Count dataset as a proxy for ridership, which is provided freely by Transport for London (TfL). The dataset describes the average number of *entrances* at each station in the Underground Network of Greater London, represented as a time series spanning 24 hours. The time series are aggregated at 15 minute intervals, resulting in 96 data points per station. This represents an average of all days in the month of November 2017, separated into weekdays and weekends. We verify from the TfL datasets that November 2017 illustrates a typical sample of winter travelling behaviour throughout the year and has been adjusted by TfL for any disruptions in the Underground service. The time series of entrances (or exits) at stations represents an aggregation of many different commuting patterns. Since the vast majority of commuting takes place on weekdays, we ignore the ridership

patterns on weekends, corresponding to Saturday and Sunday, the designated weekend in London work district. We re-factor the wide dataset of TfL arranged as aggregated counts of entrances into a longitudinal set amounting to a corpus of ≈ 4.56 million geolocated traces.

Disentangling transport demand into time geography. To characterise the nature of urban demand for public transport in cities, we decompose the diurnal and multimodel transit ridership into its characteristic unimodal components (mixtures) across a day that is learned through the use of GMM on the entrance data. Figure 1A shows a range of different demand categories represented by subpopulations that are automatically identified based on transit ridership (see Fig 1A). Each mixture has a component weight ϕ_k, with the constraint that $\sum^k \phi_i = 1$ such that the probability distribution function normalises to 1 ($k = 6$, see Fig 1). By analysing the weighted mixtures in a ternary plot, we observe that stations in Greater London display only a limited range of usage types (see Fig 1B).

Clustering of stations in space. Using a robust clustering technique over the mixtures, we classify six individual ridership patterns that naturally serve a mix of population types. Figure 1C illustrates that Greater London is spatially divided into concentric zones of development, displaying a variety of central business districts (CBD), secondary hubs (polycentric), and arterial flows of distinct linkage types (outer and inner residential, mixed-commuter, and potential feeder zones from outside the city). We reckon this space-time geography will be different for other observed cities with different patterns of concentric or polycentric development.

References

1. Dawkins, C., Moeckel, R.: Transit-Induced Gentrification: Who Will Stay, and Who Will Go? Housing Policy Debate 26(4-5), 801–818 (Sep 2016)
2. Zhang, Y., Marshall, S., Manley, E.: Network criticality and the node-place-design model: Classifying metro station areas in Greater London. Journal of Transport Geography 79, 102485 (Jul 2019)
3. Dieleman, F., Wegener, M.: Compact City and Urban Sprawl. Built Environment (1978-) 30(4), 308–323 (2004)
4. Louail, T., Lenormand, M., Picornell, M., Cant, O.G., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M.: Uncovering the spatial structure of mobility networks. Nature Communications 6 (2015)
5. Raudenbush, S.W., Sampson, R.J.: Ecometrics: Toward a Science of Assessing Ecological Settings, With Application to the Systematic Social Observation of Neighborhoods. Sociological Methodology 29(1), 1–41 (1999)
6. Taylor, B.D., Miller, D., Iseki, H., Fink, C.: Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. Transportation Research Part A: Policy and Practice 43(1), 60–77 (Jan 2009)
7. Louail, T., Lenormand, M., Cantu Ros, O.G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M.: From mobile phone data to the spatial structure of cities. Scientific Reports 4(1), 5276 (May 2014)

Comparative analysis of legal citation networks with detailed node and link properties

Joseph Hickey[1,2] and Jörn Davidsen[1,3]

[1] Complexity Science Group, Department of Physics and Astronomy, University of Calgary, Calgary, Alberta, Canada, T2N 1N4
joseph.hickey@ucalgary.ca,
[2] Bank of Canada, Ottawa, Ontario, Canada, K1A 0G9
[3] Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada, T2N 4N1

1 Introduction

Citation networks provide an intriguing means of studying human activity in fields such as science, law, and patenting [1]. While previous studies have investigated structural features of these networks, they have been limited by a lack of information about the properties of the nodes and links. As such, it is not yet well known what structural features may be revealed by information such as the level of nodes in an explicit hierarchy of journals or courts, and the sign of links. In particular, the transitive reduction link removal operation has been proposed as a way to pare a citation network down to its most essential or "informative" links [2]. This is based on a mechanism that has been proposed in some models of the growth of citation networks, in which authors make citations in one of two ways: first, by searching for relevant articles and then reading them; and second, by copying citations from documents that they have found via the first method (without reading them) [3]. We examine and compare the structural properties of legal citation networks with detailed node and link properties from three different areas of law. Despite strong differences in the three areas of law that are confirmed by an analysis of the legal topics associated with each node, the high-level properties of the three networks are very similar. However, using the information in our dataset about the "treatment" of the links (whether the citation is positive or neutral) and the hierarchical level of the court that issued the judgment, we also find evidence that the transitive reduction operation removes key structure from these networks, contrary to the proposal of Clough et al. [2]. These findings indicate that link copying is not a relevant mechanism in the growth of legal citation networks.

2 Data

We study citation networks of court decisions constructed from unique datasets in three distinct areas of Canadian law: defamation, bankruptcy, and family law. In each network, the nodes represent judgments made by courts from all levels of the court hierarchy (provincial trial-level, provincial appellate-level, and Supreme Court of Canada), and the links represent citations from newer to older judgments. In addition to the level of court, each node has one or more legal topics associated with it. These topics indicate legal issues involved in the decision. For example, topics in family law include spousal

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

support, division of family property, custody and access of children, and other topics. Additionally, each link has a "treatment" value indicating how the citing judge cited the past decision. The treatment of a link can be positive ("Followed" (F)), neutral ("Considered" or "Referred to" (CR)) or negative ("Distinguished" (D) or "Not Followed").

3 Results and Discussion

Analysis of the legal topics confirms fundamental differences in the three areas of law. Specifically, bankruptcy law consists of several strongly isolated ("siloed") sub-areas, where each sub-area pertains to a different legal topic and is readily identified as a cluster of nodes by the Infomap algorithm [4]. In contrast, defamation judgments often share the same legal topics, and clustering is dominated by temporal rather than topical structure [5]. Family law is more complex, containing isolated sub-areas such as cases involving child protection agencies, but also containing a large number of judgments with multiple topics pertaining to division of finances and parental custody of children.

However, despite these fundamental differences in the three areas of law, we show that the high-level properties of the three citation networks are very similar. In particular, and similar to results found in scientific citation networks [6], in-degree distributions are plausibly fit by a power-law with exponent between 2 and 3, directed degree-degree correlations are close to zero, and distribution of 3-node motifs follows a similar pattern in each of the three networks. Examining these properties when only certain types of links are retained in the network (e.g. keeping only the positive or neutral links) exposes key structural features of the networks, particularly in the motif analysis.

In directed acyclic graphs, such as our citation networks, only four different 3-node motifs are possible (shown on the x-axis of Fig. 1). As seen in Fig. 1, the distribution of 3-node motifs following transitive reduction resembles the distribution obtained when only the least significant, neutral links are retained in the network; conversely, when only the positive links are retained, the motif distribution is significantly different. Additionally, retaining only the positive links results in a much higher proportion of target nodes (the cited node at the end of a link) issued by the highest level of the court hierarchy, as compared to the full network, whereas considering only the links remaining after transitive reduction results in a much higher proportion of target nodes from the lowest level of the court hierarchy. Fig. 2 is an example showing why these strong differences between positive and neutral or post-transitive reduction links occur: newer judgments often make positive citations directly to highly-cited past judgments, despite the existence of alternative citation paths of neutral links that pass through intermediate judgments. These findings suggest that link copying is not relevant in the growth of legal citation networks, and suggest alternative mechanisms related to node fitness (propensity of a node to attract links, e.g. due to its hierarchical level) and link type.

References

1. Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., Stanley, H.: The science of science: From the perspective of complex systems. Phys. Rep., 714-715:1–73 (2017)
2. Clough, J.R., Gollings, J., Loach, T.V., Evans, T.S.: Transitive reduction of citation networks. J. Complex Net. 3(2), 189–203 (2015)

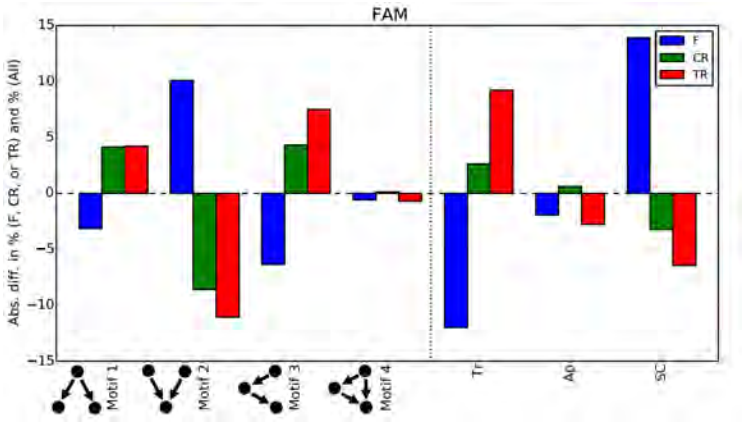


Fig. 1. Comparison of motif occurrence rates and court level of target nodes in the family law citation network. To the left of the vertical dashed line, the *y*-axis shows the difference in i) the percentage of each 3-node motif when considering only the positive (F), neutral (CR), or post-transitive reduction (TR) links, minus ii) the percentage of each 3-node motif in the full network (considering all links). To the right of the vertical dashed line, the *y*-axis shows the difference in proportion of target nodes issued by the trial (Tr), appellate (Ap) or supreme (SC) courts.

Fig. 2. Portion of a citation network showing the link patterns leading to the motif distributions shown in Fig. 1. The yellow node represents an appellate-court judgment (*Smith*), and the rest of the nodes and links constitute the tree reachable from the *Smith* in the reversed citation network. The green node is an example of a judgment that directly cites *Smith* using a positive link, while also being part of alternate paths to *Smith*. Link treatments: blue: F; grey: CR; orange: D.

3. Golosovsky, M., Solomon, S.: Growing complex network of citations of scientific papers: Modeling and measurements. Phys. Rev. E, 95(1):1–19 (2017)
4. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci., 105(4):1118–1123 (2008)
5. Leicht, E.A., Clarkson, G. Shedden, K., Newman, M.E.J.: Large-scale structure of time evolving citation networks. Eur. Phys. J. B, 59:75–83 (2007)
6. Šubelj, L., Fiala, D., Bajec, M.: Network-based statistical comparison of citation topology of bibliographic databases. Sci. Rep., 4:6496 (2014)

The Pólya filter: A parametric approach to backbone extraction in complex weighted networks

Riccardo Marcaccioli[1] and Giacomo Livan[1,2]

[1] Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6EA, UK

[2] Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, UK

1 Introduction

Understanding which nodes and links represent a set of structurally relevant interactions is of crucial importance to obtain parsimonious representations of large network datasets, often referred to as *network backbones*. Indeed, filtering out noise in order to extract meaningful backbones has shed light on the functioning of complex weighted networks of repeated interactions in a variety of disciplines, ranging from Biology [1] to Finance [2].

Earlier approaches to network backbone extraction relied on simple weight thresholding. This, however, often amounts to neglecting the multiscale nature of most real-world networks. Most methodologies put forward in recent years, instead, take such multiscale nature into account by assigning a p-value to each link in a weighted network by measuring the probability of observing its weight under a null hypothesis of partially random interactions. One of the most successful - and widely adopted - options in the literature is the disparity filter [3], which relies on a null hypothesis of uniform distribution of a node's strength over its links.

The disparity filter and other options in the same spirit (see, e.g., [4]) provide top-down approaches based on well defined null hypotheses, against which all links in a network are tested individually. While this certainly presents advantages in terms of convenience, at the same time it can lead to a lack of flexibility, as different networks may display different levels of heterogeneity, to which a 'one-fits-all' null hypothesis cannot adapt. Furthermore, most of the above filters are based on null hypotheses of partially random interactions. Yet, interactions in most natural and social systems are far from being random, as past activity naturally breeds further activity.

Here, we propose a filtering methodology based on a null hypothesis designed to respond to the specific heterogeneity of a network. We do so by contrasting links against a null hypothesis based on the Pólya urn, a well known combinatorial problem driven by a self-reinforcement mechanism according to which the observation of a certain event increases the probability of further observing it. Such a mechanism is governed by a single parameter $a > 0$ (which in the urn analogy quantifies the number of balls of a certain color added to the urn after extracting one ball of the same color), which allows to tune the null hypothesis' tolerance to a specific network's heterogeneity, and to study a continuous *family* of network backbones $\mathscr{P}_a$.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

The full description of the Pólya filter and some of its possible applications have been recently published in [5]. The paper's main results are summarised in the following.

1. We analytically demonstrate that the p-value assigned by the Pólya filter to any given link is - with excellent approximation - a function of the ratio $r = wk/s$, where w is the weight of the link being tested, k is the degree of one of the two nodes it is attached to (links can be tested with respect to both their directions), and s is its strength. Therefore, whether a link is statistically validated against the null hypothesis - and therefore included in a backbone $\mathscr{P}_a$ - is determined by an interplay between weights and topology, which accounts for the Pólya filter's ability to validate links at all scales.
2. We prove that the disparity filter [3] is recovered as a special case of the Pólya filter for $a = 1$.
3. The Pólya filter becomes increasingly restrictive with the parameter a, with fewer links being identified as statistically significant as a increases. As a result, backbones associated with higher values of a are subsets of those obtained for smaller values (see Fig. 1), i.e., $\mathscr{P}_{a_2} \subset \mathscr{P}_{a_1}$ for $a_2 > a_1$. Furthermore, we show that as backbones become more sparse with higher values of a, they retain links with higher *salience* (a recently proposed measure of link importance, which can be loosely defined as the fraction of weighted shortest-path trees a link participates in [6]) in the network.
4. We compare the Pólya filter's performance on a number of network datasets against that of 5 other well established alternatives in the literature, demonstrating its unique ability to generate backbones that are simultaneously sparse, salient, and heterogeneous.
5. We provide a criterion to filter a network against a null hypothesis tailored around its *own* heterogeneity. This is done by identifying the Pólya process whose self-reinforcement mechanism is the most likely to generate the specific network under study. Effectively, this amounts to identifying the value of a corresponding to the 'nullest' model in the Pólya family or, in other words, the Pólya process that best captures the heterogeneity of the network under consideration.

We showcase the above properties via two main applications devoted, respectively, to the US air transport network (see Fig. 1) and to the input-output network of global trade. In the latter case, we also provide evidence that links validated by the Pólya filter are highly predictive of future interactions.

Summary. We propose a novel technique to extract backbones of statistically relevant interactions between pairs of nodes in a weighted network based on the Pólya urn model. The link selection criterion underpinning the Pólya filter is based on the interplay between topology and the local relative importance of a link. This, in turn, guarantees that the filter does not perform a naive link selection merely based on retaining high strength links connecting hubs, but instead ensures a non-trivial scanning of all the relevant scales of a network. As a result, the Pólya filter generates network backbones that are simultaneously sparse, salient, and heterogeneous.

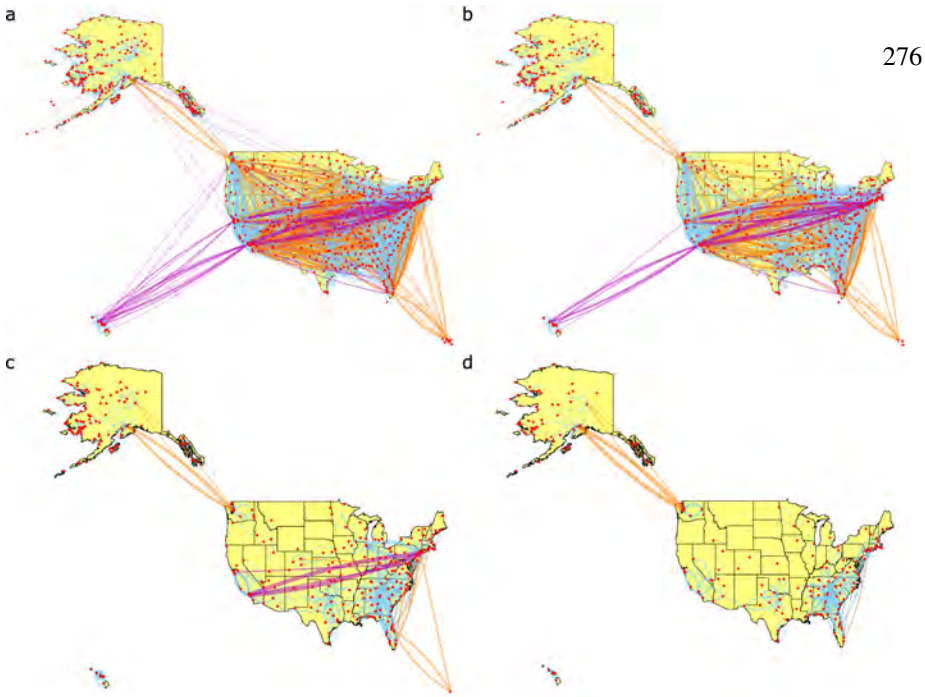


Fig. 1. Pólya backbones of the US Airports network for different values of the filter's parameter a. Links in blue, orange, and purple correspond, respectively, to short, medium, and long-haul flights according to the US Bureau of Transportation's classification. (**a**) Backbone for $a = 0.4$. (**b**) Backbone for $a = 1$, which approximately corresponds to the one obtained via the disparity filter. (**c**) Backbone for $a = 2.6$, which is the highest value of the filter's parameter where a long-haul flight (New York - Los Angeles) is retained. (**d**) Backbone for $a = 4.5$, which corresponds to the network's optimal value mentioned in point 4 of the list above. As it can be seen, higher values of a lead to sparser backbones. When tuning the Pólya filter to the network's own heterogeneity (panel (**d**)) all major long-haul flights between hubs (i.e., the links that mostly characterise the network's heterogeneity) are filtered out, resulting in a backbone of mostly regional and short-haul flights connecting airports that are often of secondary importance on the national scale. Yet, these flights provide vital connections, carrying very large numbers of passengers relative to the overall heterogeneity of the broader transport system they are embedded in.

References

1. Zhou, X., Menche, J., Barabási, A.–L., Sharma, A.: Human symptoms–disease network. Nat. Commun. 5, 4212 (2014).
2. Pozzi, F., Di Matteo, T., Aste, T.: Spread of risk across financial markets: better to invest in the peripheries. Sci. Rep. 3, 1665 (2013).
3. Serrano, M. Á., Vespignani, A., Boguná.: M. Extracting the multiscale backbone of complex weighted networks. Proc. Natl. Sci. Acad. USA 106, 6483–6488 (2009).
4. Tumminello, M., Miccichè, S., Lillo, F., Piilo, J., Mantegna, R. N.: Statistically validated networks in bipartite complex systems. PLoS ONE 6, e17994 (2011).
5. Marcaccioli, R., Livan, G.: A Pólya urn approach to information filtering in complex networks. Nat. Commun. (2019)
6. Grady, D., Thiemann, C., Brockmann, D.: Robust classification of salient links in complex networks. Nat. Commun. 3, 864 (2012).

Detecting core-periphery structures by surprise

Jeroen van Lidth de Jeude[1], Guido Caldarelli[1], and Tiziano Squartini[1]

IMT School for Advanced Studies Lucca, P.zza San Francesco 19, 55100 Lucca (Italy)
tiziano.squartini@imtlucca.it

1 Introduction

Detecting the presence of mesoscale structures in complex networks is of primary importance [1, 2]. This is especially true for financial networks, whose structural organization deeply affects their resilience to shocks propagation, node failures, etc. [3–6]. Several methods have been proposed, so far, to detect communities, i.e., groups of nodes whose "internal" connectivity is significantly large. Communities, however, do not represent the only kind of mesoscale structures characterizing real-world networks: other examples are provided by bow-tie, core-periphery and bipartite structures. In what follows, we will focus on the last two types of topological structures.

In recent years the detection of mesoscale structures has been faced by adopting a bottom-up approach, i.e., by defining a benchmark model against which to compare the actual network structure: in [7] the authors aim at identifying the most likely generative model that may have produced a given partition; in [8, 9] the authors compare the likelihood values of a stochastic block model tuned to reproduce either a core-periphery or a bipartite structure; similarly, in [10] the authors adopt a random graph model to find multiple core-periphery pairs in networks and in [11] the same authors employ the configuration model as a benchmark, showing that a single core-periphery structure can never be significant under it.

2 Results

We contribute to this stream of research by proposing a novel method to detect statistically significant bimodular structures (i.e., either bipartite or core-periphery ones). To this aim, we build upon the results of the papers [12–14] and on the very last comment that can be found in [15], by adopting a surprise-like score function. Our choice is dictated by the versatility of this kind of quantity (originally introduced to detect communities) that allows us to consider undirected as well as directed (binary) networks, a desirable feature that many of the aforementioned algorithms do not have.

Whenever community detection is carried out by maximizing the surprise, links are understood as belonging to two different categories, i.e., the *internal* ones (the ones *within* clusters) and the *external* ones (the ones *between* clusters). On the other hand, whenever one is interested in detecting bimodular structures (be they bipartite or core-periphery), three different "species" of links are needed (e.g., core, core-periphery and periphery links). This is the reason why we need to consider the multinomial version of the surprise, whose definition reads

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

$$S_{\parallel} \equiv \sum_{i \geq l_c^*} \sum_{j \geq l_{cp}^*} \frac{\binom{V_c}{i}\binom{V_{cp}}{j}\binom{V-(V_c+V_{cp})}{L-(i+j)}}{\binom{V}{L}} \tag{1}$$

and that we will refer to as to the *bimodular surprise* (V is the total number of node pairs; V_c is the number of core nodes pairs; V_{cp} is the number of node pairs between the core and the periphery; L is the total number of links; l_c^* is the number of core links; l_{cp}^* is the number of links between the core and the periphery). The presence of three different binomial coefficients allows three different kinds of links to be accounted for. From a technical point of view, $S_{\parallel}$ is a p-value computed on a multivariate hypergeometric distribution.

Let us first calculate $S_{\parallel}$ for a bipartite network defined by the values of parameters $V_c = \frac{N_1(N_1-1)}{2}$ (here, the label c indicates the internal volume of one of the two layers), $V_{cp} = N_1 N_2$, $l_c^* = 0$ and $l_{cp}^* = L$ (l_{cp}^* is the number of links between the two layers, coinciding with the total number of links, in our example). The latter condition implies that only the addendum corresponding to $i = 0$, $j = l_{cp}^* = L$ survives; thus, our bimodular surprise reads

$$S_{\parallel} = \frac{\binom{V_{cp}}{l_{cp}^*}}{\binom{V}{l_{cp}^*}} = \frac{\binom{N_1 N_2}{l_{cp}^*}}{\binom{(N_1+N_2)(N_1+N_2-1)/2}{l_{cp}^*}} \tag{2}$$

which *can* be significant, as it should be: in fact, a number of inter-layer links exists above which the observed bipartite structure is significantly denser than its random counterpart.

Let us now move to analyze some real-world systems: we will employ our novel definition of surprise to understand if the considered networks have a significant bimodular structure (i.e. either bipartite or core-periphery). To this aim, we will search for the (optimal) partition that minimizes $S_{\parallel}$ by employing a modified version of the PACO algorithm [14] a Python version of which is freely available at [16].

As a first example, we consider the social network showing the relationships between US political blogs. Any two blogs are linked if one of the two references the other. As shown in 1, a core of the most influential blogs (be they republican or democratic), surrounded by a periphery of loosely connected, less important blogs is clearly visible. Differently from the community structure that shows republican blogs and democratic blogs as belonging to different groups, our core-periphery structure highlights a different organizing principle, based on the blogs overall importance irrespectively of their political orientation.

As a second example, let us consider the US airports network in its directed representation. Examples of core airports are the ones of New York, Indianapolis, Salt Lake City, Seattle, etc. The periphery airports are preferentially attached to the core ones. This system shares interesting similarities with the NetSci co-authorship network (not shown here - see [17]): each core airport, in fact, seems to be surrounded by a quite large number of periphery airports, sharing few internal connections.

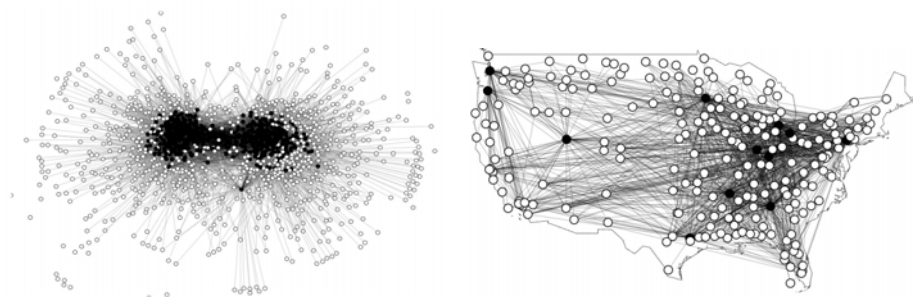


Fig. 1. Left panel: core-periphery structure of US political blogs: a core of the most influential blogs (be they republican or democratic), surrounded by a periphery of loosely connected, less important blogs is clearly visible. Notice that blogs are grouped independently from their political orientation. Right panel: core-periphery structure of US airports.

Summary. Detecting the presence of mesoscale structures in complex networks is of primary importance. This is especially true for financial networks, whose structural organization deeply affects their resilience to events like default cascades, shocks propagation, etc. Several methods have been proposed, so far, to detect communities, i.e., groups of nodes whose internal connectivity is significantly large. Communities, however do not represent the only kind of mesoscale structures characterizing real-world networks: other examples are provided by bow-tie structures, core-periphery structures and bipartite structures. Here we propose a novel method to detect statistically significant bimodular structures, i.e., either bipartite or core-periphery ones. It is based on a modification of the surprise, recently proposed for detecting communities. Our variant allows for bimodular nodes partitions to be revealed, by letting links to be placed either 1) within the core part and between the core and the periphery parts or 2) between the layers of a bipartite network. From a technical point of view, this is achieved by employing a multinomial hypergeometric distribution instead of the traditional (binomial) hypergeometric one.

References

1. Fortunato S. and Hric D., *Phys. Rep.*, 659 (2016) 1.
2. Khan B. S. and Niazi M. A., *arXiv:1708.00977* (2017).
3. Craig B. and von Peter G., *J. Financ. Intermediat.*, 23 (2014) 322.
4. in 't Veld D. and van Lelyveld I., *J. Bank. Finance*, 49 (2014) 27.
5. Fricke D. and Lux T., *Comput. Econom.*, 45 (2014) 352.
6. Luu D. T., Napoletano M., Barucca P. and Battiston S., *arXiv:1802.02127* (2018).
7. Zhang X., Martin T. and Newman M. E., *Phys. Rev. E*, 91 (2015) 032803.
8. Barucca P. and Lillo F., *Chaos, Solitons, Fractals, 88 (2016) 244.*
9. Barucca P. and Lillo F., *Comput. Manag. Sci.*, 15 (2018) 33.
10. Kojaku S. and Masuda N., *Phys. Rev. E*, 96 (2017) 052313.
11. Kojaku S. and Masuda N., *New J. Phys.*, 20 (2018) 043012.
12. Aldecoa R. and Marin I., *Sci. Rep.*, 3 (2013) 2216.
13. Aldecoa R. and Marin I., *Sci. Rep.*, 3 (2013) 1060.

14. Nicolini C. and Bifone A., *Sci. Rep.*, 6 (2016) 19250.
15. Traag V. A., Aldecoa R. and Delvenne J. C., *Phys. Rev. E*, 92 (2015) 022816.
16. https://github.com/jeroenvldj/bimodular_surprise
17. van Lidth de Jeude J., Caldarelli G., Squartini T., *Europhys. Lett.*, 125 (2019) 68001.

A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks

Federica Parisi[1], Tiziano Squartini[1], and Diego Garlaschelli[1]

IMT School for Advanced Studies Lucca, P.zza San Francesco 19, 55100 Lucca (Italy)
tiziano.squartini@imtlucca.it

1 Introduction

Network reconstruction is an active field of research within the broader field of complex networks [1, 2]. Among the methods proposed so far, some assume that the constraints concerning the binary and the weighted network structure jointly determine the reconstruction output [3] while others consider the weights estimation step as completely unrelated to the binary one [4, 5]. Amidst the former ones, a special mention is deserved by the Enhanced Configuration Model (ECM) [3]; examples of algorithms belonging to the second group are those iteratively adjusting the link weights on top of some previously-determined topological structure (e.g. via the recipe firstly proposed in [6]), in such a way to satisfy the constraints concerning strengths *a posteriori*. This kind of procedure, however, assigns weights deterministically, thus being unable to provide confidence bounds accompanying the weight estimates [7] and basically giving zero probability to any real-world network. Two-step algorithms also exist [8] that attempt to overcome the lack of binary information for the estimation of topology: however, they are characterized by a high computational complexity and only heuristically motivated.

2 Results

Available reconstruction algorithms implement and combine in different fashions the two estimation steps discussed above. In order to unambigouously asses the goodness of a given method, we employ the likelihood $\ln Q(\mathbf{W}^*)$ as a score function, aimed at quantifying how likely the structure of a given real-world network $\mathbf{W}^*$ is reproduced by a given algorithm. As shown in table 1 (see also [9]), the probability of reproducing a real-world structure is either zero (further impliying that $\ln Q(\mathbf{W}^*) = -\infty$) or rapidly vanishing as the number of nodes grows.

Here we develop a theoretical framework that provides an analytical, unbiased procedure to estimate the weighted structure of a network, once its topology has been determined, thus extending the Exponential Random Graph (ERG) recipe. In our approach, information about the topological structure (either available *ab initio* or obtained by using any of the existing algorithms) is treated as *prior* information; together with the proper weighted constraints, it represents the input of our generalized reconstruction procedure. The probability distribution describing link weights is, then, determined by maximizing the key quantity of our algorithm, i.e. the *conditional entropy*

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Method	Topology	Weights	Log-likelihood
MaxEnt	D	D	$-\infty$
Minimum-Density	D	D	$-\infty$
Copula approach	P	D	$-\infty$
Drehmann & Tarashev	P	D	$-\infty$
Montagna & Lux	P	D	$-\infty$
Mastromatteo et al.	P	D	$-\infty$
Gandy & Veraart	P	D	$-\infty$
dcGM	P	D	$-\infty$
MECAPM	P	P ($w_{ij} \in \mathbb{N}$)	$-\infty$
fitness-induced DECM	P	P ($w_{ij} \in \mathbb{N}$)	$-\infty$
Hałaj & Kok	P	P	$\in \mathbb{R}$
Moussa & Cont	P	P	$\in \mathbb{R}$

Table 1. Overview of the reconstruction methods reviewed in [1]. The letter "P" indicates that the considered estimation step is probabilistic in nature while the letter "D" indicates that it is deterministic (see [9] for the corresponding references).

$$S(\mathscr{W}|\mathscr{A}) = -\sum_{\mathbf{A}\in\mathbb{A}} P(\mathbf{A}) \int_{\mathbb{W}_{\mathbf{A}}} Q(\mathbf{W}|\mathbf{A}) \log Q(\mathbf{W}|\mathbf{A}) d\mathbf{W} \tag{1}$$

under a properly-defined set of constraints. This algorithm returns a *conditional probability distribution* depending on a vector of unknown parameters (say $\boldsymbol{\lambda}$); in alignment with previous results, their estimation is carried out by considering the *generalized likelihood*

$$\mathscr{G}(\boldsymbol{\lambda}) = -H_{\boldsymbol{\lambda}}(\mathbf{W}^*) - \sum_{\mathbf{A}\in\mathbb{A}} P(\mathbf{A}) \log Z_{\mathbf{A},\boldsymbol{\lambda}} \tag{2}$$

(with $\mathbf{W}^*$ representing the observed configuration). In our work, we explore two possible specifications of the framework above: the simplest - and still most powerful - one (i.e. the CReM$_B$ model) is defined by the conditional, pair-specific weight distribution

$$q_{ij}(w_{ij}|a_{ij}=1) = \lambda_{ij} e^{-\lambda_{ij} w_{ij}},\ w_{ij} > 0 \tag{3}$$

whose tensor $\lambda_{ij}\ (\forall\, i \neq j)$ of unknown parameters is determined by solving the set of $O(N^2)$ *decoupled* equations $\langle w_{ij} \rangle = \frac{f_{ij}}{\lambda_{ij}} = w^*_{ij}\ (\forall\, i \neq j)$. In fig. 1 we compare the reconstruction accuracy of the two specifications of our framework.

Summary. The knowledge of the structure of a financial network gives valuable information for estimating the systemic risk. However, since financial data are typically subject to confidentiality, network reconstruction techniques become necessary to infer both the presence of connections and their intensity. Recently, several "horse races" have been conducted to compare the performance of these methods. These comparisons were based on arbitrarily chosen metrics of network similarity. Here we establish a generalised likelihood approach to rigorously define and compare methods for reconstructing weighted networks. The crucial ingredient is the possibility to input any purely

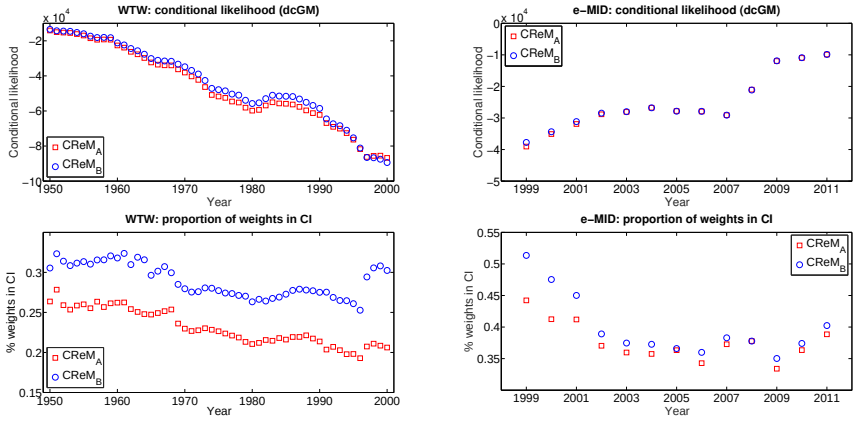


Fig. 1. Top panels: comparison between the conditional likelihood function of the CReM$_A$ and the CReM$_B$ model (red squares and blue circles, respectively), for the WTW and e-MID. The reconstruction accuracy obtainable by employing the CReM$_B$ model is comparable with the one obtainable by employing the CReM$_A$ model; still, it is achievable with much less computational effort. Bottom panels: percentage of observed weights that fall into the CI around their estimate.

binary reconstruction method and, conditionally on it, to exploit aggregate information about link weights in order to unbiasedly reconstruct a weighted network. Our results indicate that the best method is obtained by "dressing" the best-performing, available binary method with an exponential distribution of weights. The method is fast, unbiased and reproduces empirical networks with highest generalised likelihood.

References

1. Squartini, T. et al. Reconstruction methods for networks: the case of economic and financial systems, *Physics Reports* 757, 1-47, doi: 10.1016/j.physrep.2018.06.008 (2018).
2. Cimini, G. et al. The statistical physics of real-world networks, *Nature Reviews Physics* 1(1), 58-71, doi: 10.1038/s42254-018-0002-6 (2019).
3. Mastrandrea, R. et al. Enhanced reconstruction of weighted networks from strengths and degrees, *New Journal of Physics*, 16(4) (2014).
4. Andrecut, M. Systemic risk, maximum entropy and interbank contagion, *International Journal of Modern Physics C*, 27(12) (2016).
5. Hałaj, G., Kok, C. Assessing interbank contagion using simulated networks, *Computational Management Science*, 10(2-3) (2013).
6. Bacharach, M. Estimating nonnegative matrices from marginal data, *International Economic Review*, 6(3) (1965).
7. Gandy, A., Veraart, L. A Bayesian methodology for systemic risk assessment in financial networks, *Management Science*, 63(12) (2016).
8. Cimini, G. et al. Estimating topological properties of weights networks from limited information, *Physical Review E*, 92(4), 040802 (2015).
9. Parisi, F., Squartini, T., Garlaschelli, D. A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks, *arXiv:1811.09829* (2019).

Core–Periphery Structure in Directed Networks

Andrew Elliott[1,2], Angus Chiu[2], Marya Bazzi[1,3,4], Gesine Reinert[1,2] and Mihai Cucuringu[1,2,4]

[1] The Alan Turing Institute, London, UK
[2] Department of Statistics, University of Oxford, Oxford, UK
[3] Mathematical Institute, University of Warwick, Coventry, UK
[4] Mathematical Institute, University of Oxford, UK

1 Introduction

In undirected networks, core–periphery is a meso-scale network structure typically consisting of a well connected core and a periphery that is well connected to the core but sparsely connected internally [3, 15, 20]. In [8] we propose a novel generalisation of core–periphery structure to directed networks. Related approaches which can be applied to the analysis of directed networks include [2, 18], SBMs [13, 17] and the bow-tie structure [4].

This extended abstract focuses on one particular core–periphery structure consisting of four sets (two core sets and two periphery sets) which is constructed as follows. We split each of the sets in the traditional formulation (one core and one periphery) into one subset that has only incoming edges and another subset that only has outgoing edges, yielding four sets, $\mathscr{C}_{in}$ (*core-in*), $\mathscr{C}_{out}$ (*core-out*), $\mathscr{P}_{in}$ (*periphery-in*) and $\mathscr{P}_{out}$ (*periphery-out*). Within each of the two core sets ($\mathscr{C}_{in}$ and $\mathscr{C}_{out}$) and periphery sets ($\mathscr{P}_{in}$ and $\mathscr{P}_{out}$), we follow the undirected formulation and assume that the two core sets are fully internally connected, and the two periphery sets do not connect to each other and do not have internal edges. Directed connections from $\mathscr{C}_{out}$ to $\mathscr{P}_{in}$, from $\mathscr{C}_{out}$ to $\mathscr{C}_{in}$, and from $\mathscr{P}_{out}$ to $\mathscr{C}_{in}$, are permitted, leading to the structure shown in Fig. **1**A. This structure differs from bow-tie in that bow-tie is a uni-directional flow-based structure.

To detect this structure, we introduce four methods, having different trade-offs between computational complexity and accuracy. In order of speed, these are: an extension of the Low-Rank approach from [7] which is based on the singular value decomposition of the adjacency matrix combined with an in– and out–degree based score (LowRank), an extension of the hub and authority scores of the HITS algorithm to scores for periphery sets from [10] (Hits), a HITS-type score which instead of using hub and authority scores rewards similarity to the hypothesized structure (AdvHits), and maximum likelihood estimation in an appropriate stochastic block model (MaxLike), see also [9].

2 Results

We validate our proposed approaches on three synthetic benchmark network data sets, and compare their performance to the general methods SAPA [16], Di-Sum [14], Graph-Tool [11, 12] and simply the in-and out-degrees. We observe that our specialised approaches perform at least as well as the general methods with similar run times. They

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

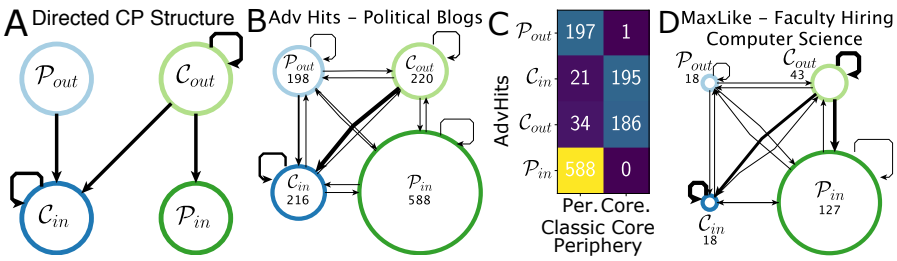


Fig. 1. A Network Diagram detailing our hypothesised '*ideal*' structure. **B** Partition of the Political Blog Data by our ADVHITS method. **C** Comparison between the undirected core periphery structure and our directed core periphery data using our ADVHITS method on Political Blog Data from [1]. **D** Partition of the Computer Science Faculty Hiring Data by our MAXLIKE method.

often outperform the general methods when the planted structure is particularly weak, highlighting the effectiveness of using specialised techniques for the proposed structure.

Next, we apply our methods to a network of political blogs from [1], and a faculty hiring network from [5]. In both cases, we test if the structures are statistically significant with respect to a directed Bernoulli random graph null model, and with respect to a directed configuration model null model. In both examples, ADVHIST and MAXLIKE yield significant differences. These methods are then used for the structure detection. For space reasons, we illustrate the results only for one method per network.

In the political blogs data set [1] (division shown in Fig. **1**B), ADVHITS gives a division of the classically undirected core into two components (see confusion table in Fig. **1**C), a $\mathscr{C}_{in}$ core and a $\mathscr{C}_{out}$ core. The $\mathscr{C}_{in}$ core turns out to contain relatively few 'blogspot' sites; these are free blogging sites which require less expertise to set up than a full website. We hypothesise that the $\mathscr{C}_{in}$ core has a relatively high proportion of authorities which are highly referenced (with high in-degree), whereas $\mathscr{C}_{out}$ core contains mainly blogs which link to a large number of other blogs (with high out-degree).

In the faculty hiring data set, our MAXLIKE approach (division shown in Fig. **1**D) uncovers the expected structure from [5] with $\mathscr{C}_{out}$ representing the top research universities and $\mathscr{P}_{in}$ representing the remaining universities hiring faculty from the top institutions. The set $\mathscr{C}_{in}$ turns out to consist of Canadian universities which have a large number of links with the top US schools, but also appear to strongly recruit from other Canadian schools in $\mathscr{C}_{in}$. This structure was not previously uncovered.

Summary. We introduce a novel directed core periphery structure, with two cores and two peripheries, and four algorithms to detect this structure, with different time quality trade-offs. All of these algorithms are able to uncover the structure in a synthetic model, and often outperform classical methods.

Our methods reveal previously unobserved structures in two real-world data sets. In the political blogs data set, we find a division of the known undirected core–periphery into what may be an authoritative core and what may be a less authoritative core. In the faculty hiring data, in addition to the structure observed in [5], we find a separate core which corresponds to Canadian Universities who attract academics from high-ranking American institutions, but also strongly recruit from other Canadian schools.

Thus, the directed core–periphery structure is a structure which is able to reveal interesting features in real data sets and thus warrants further investigation, including more detailed comparison with other approaches.

References

1. Adamic LA, Glance N. 2005 The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* pp. 36–43. ACM.
2. Beguerisse-Díaz M, Garduno-Hernández G, Vangelov B, Yaliraki SN, Barahona M. 2014 Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *Journal of the Royal Society Interface* **11**, 20140940.
3. Borgatti SP, Everett MG. 1999 Models of core/periphery structures. *Social Networks* **21**, 375–395.
4. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. 2000 Graph structure in the Web. *Computer Networks* pp. 309–320.
5. Clauset A, Arbesman S, Larremore DB. 2015 Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* **1**, e1400005.
6. Csermely P, London A, Wu LY, Uzzi B. 2013 Structure and dynamics of core/periphery networks. *J. Complex Networks* **1**, 93–123.
7. Cucuringu M, Rombach P, Lee SH, Porter MA. 2016 Detection of core–periphery structure in networks using spectral methods and geodesic paths. *European Journal of Applied Mathematics* **27**, 846–887.
8. Elliott, A., Chiu, A., Bazzi, M., Reinert, G., Cucuringu, M. (2019). Core–Periphery Structure in Directed Networks. In preparation.
9. Karrer B, Newman ME. 2011 Stochastic blockmodels and community structure in networks. *Physical Review. E* **83**, 016107.
10. Kleinberg JM. 1999 Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* **46**, 604–632.
11. Peixoto TP. 2014 Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review. E* **89**, 012804.
12. Peixoto TP. 2014a The graph-tool python library.
13. Peixoto TP. 2013 Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X* **4**.
14. Rohe K, Qin T, Yu B. 2016 Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences* **113**, 12679–12684.
15. Rombach P, Porter MA, Fowler JH, Mucha PJ. 2017 Core-Periphery Structure in Networks (Revisited). *SIAM Review* **59**, 619–646.
16. Satuluri V, Parthasarathy S. 2011 Symmetrizations for clustering directed graphs. In *Proceedings of the 14th International Conference on Extending Database Technology* pp. 343–354. ACM.
17. Snijders TA, Nowicki K. 1997 Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification* **14**, 75–100.
18. Travencolo, B., Viana, M., da F. Costa, L. (2009). Border Detection in Complex Networks. New Journal of Physics. 11. 10.1088/1367-2630/11/6/063019.
19. Tudisco F, Higham DJ. 2019 A nonlinear spectral method for core–periphery detection in networks. *Siam Journal of Data Science* **1**, 269–292.
20. Zhang X, Martin T, Newman ME. 2015 Identification of core-periphery structure in networks. *Physical Review. E* **91**, 032803.

Friendship Concept and Community Network Structure among Elementary School and University Students

Ana María Hernández-Hernández[1], María Dolores Viga-de Alba[2] Rodrigo Huerta-Quintanilla, Efrain Canto-Lugo, Hugo Laviada-Molina, and Fernanda Molina-Segui

[1] Centro de Investigación y de Estudios Avanzados del Institute Politécnico Nacional, Unidad Mérida
amhernandezh@gmail.com,
WWW home page: www.cinvestav.mx
[2] Escuela de Ciencias de la Salud-UNEXMAR, Universidad Marista de Mérida, Mérida, Yucatán, México

1 Introduction

Friendship, as a historical phenomenon, is complex, dynamic, and sensitive to influence [1, 2]. Social scientists have found that the concept of friendship is not as simple as we may think but depends on several variables [3–5]. Friendship, identified as a basis of human and psychological wellbeing [6, 7], is influenced by several environments in which people interact with each other, with their families, and with other social communities such as schools, churches, neighborhoods, etc. In this study, we adopt an approach that allows us to develop a context for the findings reported concerning friendship, the kind of effects that are produced across a social networks structure, and the communities defined in social network analysis [8–10]. We analyze four friendship networks, three from elementary schools and one from a university. Two of the three elementary schools are in rural areas. The first two elementary schools are public schools located in Temozón South and Abalá, Yucatán, respectively. Those are rural areas where the principal activities are commerce, agriculture, construction, and manufacturing. The third elementary school is a public school located in the urban area of Mérida, Yucatán. The university in which we conducted our study is a private one, located in Mérida, Yucatán. In this work, we are interested in identifying how communities emerge in the presence of friendships and mixed links (relatives considered also as friends), and also we were looking for differences in the communities composition and topology of the networks according to where the schools are located and the socioeconomic position of those locations. To do this we employ Newmans algorithm for all four networks. We consider this study important because it allows us to provide an analytic and computational basis to observational studies, thus providing the opportunity to engage in interdisciplinary work and open new lines of research for applied physics.

2 Results

The collection of data was done through a survey. At elementary schools, the surveys were applied on paper due to the participant's ages and the limited access to the internet,

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

especially in rural areas. For university students, the same instrument was applied using a web platform designed for that purpose. Each student individually fills the survey (under the supervision of interviewers) using a computer. Once the data were acquired, the adjacency matrix was constructed taking into account that the links should be confirmed, it means if the participant i mention participant j in the survey, the link exists only if j also mentions i. This confirmation was done to avoid bias in the study. In elementary schools, relatives were also considered as friends (mixed links) by the students. The three elementary school networks are all connected if the mixed links are considered. If we do not consider these links, the networks corresponding to schools 2 and 3 become disconnected. On the other hand, the university network is disconnected and is formed by 61 components. We calculate the communities from the largest component of the network. We start the network analysis by calculating main properties such as the average degree k, the clustering coefficient C, and density ρ. We observed that elementary schools networks have large values of k compared with the university network. One interesting thing about the rural schools is that they have a large number of mixed links because of the siblings and relatives that students have on school, and also consider being friends. It is a large difference between the two rural schools and the urban school in terms of the mixed links. To find the community structure of the networks, we used Newmans algorithm [11, 12] and investigated the relationship between the friendship, family ties and that structure. On the urban elementary school, eight of the twenty-five communities found were similar to classrooms composition (the initial grades of scholarly). In contrast, the communities found in rural areas were a mix of students from different classrooms. We also considered the networks and the communities obtained where mixed links are not considered (figure 1). Once the mixed links were removed some communities start to be closer to the classrooms. Mixed links are the ones that help to establish a connection between classrooms and breaks in a way the spatial confinement. For this reason, when the mixed links are removed, some communities show similar composition to classrooms.

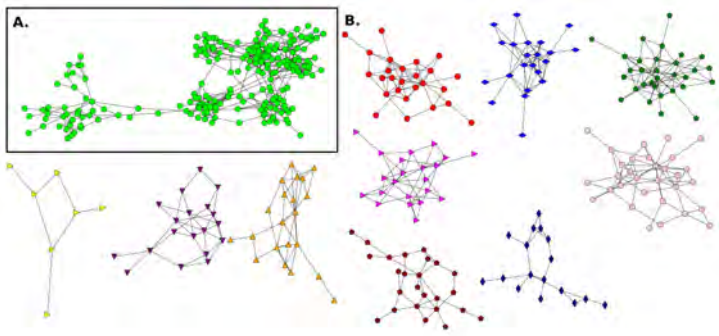


Fig. 1. Giant component of Elementary School network and its communities without mixed links E2(NF). A. The giant component has n_g = 221 (nodes), m_g = 536 (links) and k_g = 4.85. **B.** Communities detected in the giant component.(doi:10.1371/journal.pone.0164886.g006).

Therefore, according to the results, spatial confinement favors the formation of friends at elementary school. Most of students' friend (in elementary schools) are from the same classroom. With university students, it is a different situation. These students have a more defined concept of friendship and therefore do not necessarily consider classmates as friends. One may conclude that spatial confinement has no relevance to having a friend at the university. Even though we also find communities that can be large, most students are in different classrooms. At this level, we do not have mixed links. One of the limitations of this type of work is the difficulty of obtaining the data, which is a lot of work and is a slow process that does not provide all the information we need. Another limitation is not having all the information about external factors, which implies the need for not only working with students but also with their families. This kind of work is important to know the network structure of scholar networks for studying dispersion diseases in the future, especially in a zone like Yucatán (which has tropical weather and it is a high incidence of diseases like dengue) the study of this structures can be useful.

Summary. We analyzed the structure and the communities of four friendship networks and found significant differences among elementary schools and university networks. In elementary schools, the students make friends mainly in the same classroom, but there are also links among different classrooms because of the presence of siblings and relatives in the schools. Once the links between siblings and relatives are removed, the communities resembled the classroom composition.

References

1. Matthews SH.: Friendships through the life course: Oral biographies in old age. Sage Publications (1986)
2. Adam R.G., Blieszner R., de Vries B.: Age, gender and study location effects on definitions of friendship in the third age. J. Aging Stud. 1(14), 117–133 (2000)
3. Sherman A.M., de Vries B., Lansford J.E.: Friendship in childhood and adulthood: Lessons across the life span. Int. J. Aging Hum. Dev. 51(1), 31–51 (2000)
4. Hartup W.W., Stevens N.: Friendship and adaptation across the life span. Curr. Dir. Psychol. Sci. 8(3), 76–79 (1999)
5. Dunn J., McGuire S.: Sibling and peer relationships in childhood. J. Child. Psychol. Psychiat. 33(1), 67–105 (1992)
6. Snyder C.R., Lopez S.J.: Handbook of positive psychology. Oxford University Press (2002)
7. Harvey J.H., Pauwels B.G., Zickmund S.: Relationship connection: The role of minding in the enhancement of closeness. Oxford University Press. 423–433 (2005)
8. Vega-Redondo F.: Complex social networks. Cambridge University Press (2007)
9. Souma W., Fujiwara Y., Aoyama H.: Complex networks and economics, Physica A. 324(1-2), 396–401 (2003)
10. Boccaletti S., Latora V., Moreno Y., Chavez M., Hwang D.U.: Complex networks: Structure and dynamics. Phys. Rep. 424(4-5), 175–308 (2006)
11. Newman M.E.J.: Modularity and community structure in networks. PNAS. 103, 8577–8582 (2006)
12. Newman M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E. 3(74), 036104 1 – 19 (2006)

PageRank extremes and local dependence for random graph

Maxim S. Ryzhov[1] and Natalia M. Markovich[1]

V.A. Trapeznikov Institute of Control Sciences Russian Academy of Sciences
Profsoyuznaya Str. 65, 117997 Moscow,Russia
ryzhov@phystech.edu, markovic@ipu.rssi.ru

1 Introduction

Our purpose in the current work is to determine an extremal index (EI) of the particular node for the given directed graph $G = (V,E)$. In the context of random graphs the EI metric indicates the ability of a randomly selected node to attract highly ranked nodes in its orbit. The stationary sequence $\{X_n\}_{n\geq 1}$ with the distribution function $F(x)$ is said to have EI $\theta \in [0,1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that it holds mixing condition $D(u)$ and

$$\lim_{n\to\infty} n(1-F(u_n)) = \tau, \quad \lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta}, \tag{1}$$

where $M_{k,l} = max\{X_k,\ldots,X_l\}, 0 \leq k < l, \quad M_n = M_{0,n-1}$ [1]. The EI shows relations between the distributions of extremes and a random variable [1]

$$P(M_n \leq u_n) = (F(u_n))^{n\theta} + o(1), n \to \infty. \tag{2}$$

On the first side, an assumption can be made that the node EI is a measure of the current node dependence on neighbours in the graph G. On the other hand, the consideration were given in [2] that PageRank in random graph is an autoregressive process with random coefficients and a random depth of dependence on it. By Google's definition [3] PageRank (PR) is determined as the rank $R(p_i) = R_i$ of node (Web page) p_i by

$$R(p_i) = \sum_{p_j \in N(p_i)} \frac{c}{D_j} R(p_j) + (1-c)q_i, \quad i = \overline{1,v}, \tag{3}$$

where $N(p_i)$ is the set of incoming nodes, D_j is an out-degree of node p_j, $c \in (0,1)$ is a damping factor ($c = 0.85$ as an average probability to browse a web-page connected with current one [3]), $q_i \geq 0$ is a node personalization, $v = |V|$. Thus, local graph structure can be defined as branching tree with the investigated node as a root, a kind of Thorny Branching Tree (TBT) [4], based on PageRank relations between nodes. That means the node EI value should be found with an attribute sequence $\{X_n\}_{n\geq 1}$ for nodes, belonged to the associated node TBT. The following algorithm of the node EI estimation contains description of the mention below assumptions [6, 7]:

[0]The reported study was partly funded by RFBR, project number 19-01-00090 (recipient N.M. Markovich, conceptualization, mathematical model development, methodology development; recipient M. S. Ryzhov, numerical analysis, validation

1. Estimate the PR (3) values of each graph $G = (V,E)$ node by the chosen method.
2. For the chosen node p_i receive a node sequence $\{p_{j_k} : k \geq 1, p_{j_1} = p_i\}$ associated with the sequence of the k largest values $\{\varepsilon^i_{j_k} > 0\}_{k\geq 1}$, where

$$\varepsilon^i_j = \begin{cases} 1, & p_i = p_j, \\ \frac{c^{s-1}R_j}{\prod_{m=1}^{s-1} D_{j_m} R_i}, & \exists (p_{j_1}, \ldots, p_{j_s}) = min\{(p_{j_1}, \ldots, p_{j_f}) : \\ & \forall m = \overline{2,f} \rightarrow (p_{j_{m-1}}, p_{j_m}) \in E, p_{j_1} = p_j, p_{j_s} = p_i\}, \\ 0, & otherwise. \end{cases} \tag{4}$$

3. The EI value is empirically calculated by an appropriate estimator. The commonly used estimator is the blocks estimator [5] for sequence of interesting nodes attributes

$$\widehat{\theta}_{Bl,n}(u) = \frac{n\sum_{i=1}^{l} \mathbb{I}\left(M_{(i-1)r,ir} > u\right)}{rl\sum_{i=0}^{n-1} \mathbb{I}\left(X_i > u\right)}, \tag{5}$$

   where $r = [\frac{n}{l}]$ is a block size, l is a number of blocks. According to sliding blocks model [7] if a node attends in different blocks, its copy should be added in each one (as on Fig. 1). Here a block is a group of the graph nodes having incomming edges with the same parent node. Level u is chosen with the bootstrap method [6].
4. Estimate the EI of the enlarging length $k, k+1, \ldots$ node sequence until a stable value is reached.

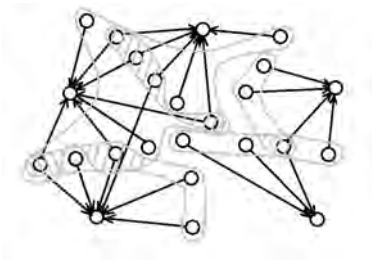


Fig. 1. Examples of the block definition by sliding blocks model.

2 Results

The EI was estimated by (5) for each node with the sequence of PageRanks (3) θ_{PR}, personalisations θ_{prs} and in-degree counts θ_{ind} provided by the local TBT. As in [7] it were received that $\theta_{PR} = 1 - \frac{1}{E(N_i)}$ or $\theta_{PR} = \frac{1}{E(N_i)}$, where $E(N_i)$ is an average node in-degree value in TBT.

$$\theta_{PR} = \begin{cases} \theta_{prs}, & if\, \alpha_{prs} \geq \alpha_{ind}, \\ \theta_{ind}, & otherwise. \end{cases} \tag{6}$$

Here α_{prs} and α_{ind} are the tail indexes of personalisations and in-degree distributions in node TBT. The latter outcome is in agreement with the results of [2] and [10], where a heaviness of PRs distribution is equal to the heaviest one from in-degree and personalisation distributions.

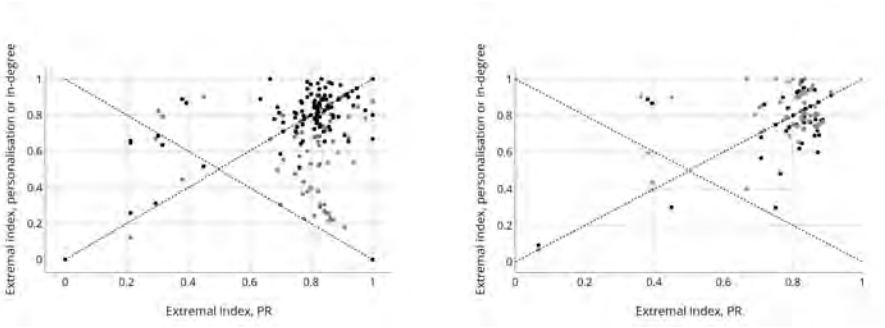


Fig. 2. Plots of the EI value for PR versus the EI value for the personalisation (grey dots) and the in-degree value (black dots) for the particular graph node. Graph is generated with Forest Fire [8] (left) and Erds-Rnyi [9] (right) models.

References

1. Leadbetter M. R., Extremes and local dependence in stationary sequences, Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete, 1983, P. 291-306.
2. Markovich N.M., Extremes in Random Graphs Models of Complex Networks, arXiv:1704.01302v1[math.ST], 5 Apr. 2017.
3. Brin S.,Page L., The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 1998, P. 107-117.
4. Chen N., Litvak N., Olvera-Cravioto M., Ranking algorithms on directed configuration networks, Memorandum of the Department of Applied Mathematics, Department of Applied Mathematics, University of Twente, 2015, V. 2046.
5. Beirlant J. ,Goegebeur Y. ,Teugels J.,Segers J., Statistics of Extremes: Theory and Applications, Wiley, Chichester, West Sussex , 2004.
6. Krieger U.,Markovich N.,Ryzhov M., Nonparametric Analysis of Extremes on Web Graphs : PageRank Versus Max-Linear Model, DCCN 2017, Distributed Computer and Communication Networks, P. 13-26.
7. Markovich N.,Ryzhov M., Random graph node classification by extremal index of PageRank, DCCN 2019, Distributed Computer and Communication Networks.
8. Leskovec J., Lang K.J., Dasgupta A., Mahoney M.W., Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters, eprint arXiv:0810.1355, 2008.
9. Erds P., Rnyi A., On Random Graphs, Publicationes Mathematicae. 6: 290297, 1959.
10. Volkovich Y.,Litvak N., On the exceedance point process for a stationary sequence, Advances in Applied Probability, 2010, P. 577-604.

A Knowledge-graph based Taxonomy Construction Method

András London[1,2], János Zsibrita[2], and Rio Fear[3]

[1] University of Szeged, Institute of Informatics, Arpád tér 2, H-6720, Szeged Hungary
london@inf.u-szeged.hu
[2] Black Swan Data, Zászló utca 4, H-6722 Szeged, Hungary
[3] Black Swan Data, 15th Floor, 10 York Rd, London SE17ND, United Kingdom

1 Introduction

A taxonomy is a hierarchically organized categorization of concepts or entities, for example a Wikipedia category, an ACM Classification System, or an Amazon Product Category. For a great many companies around the world domain-specific taxonomies form a crucial component of the provision data-driven solutions: they can help in search optimization, browsing, organization and storage of information, and much more besides. However, the creation of taxonomies is invariably a highly manual process which is time-consuming, expensive and generally unsustainable at scale, especially in fast changing domains (e.g. news and certain products), therefore an effective method of automated taxonomy generation could be highly valuable. Automated taxonomy building has been well researched in the recent years. Most approaches apply NLP tools to a text corpus e.g. [2], some of them utilize knowledge-graphs, e.g. [5], like Wikipedia or WordNet, while others combine the previous approaches e.g. [3].

In this work we provide a simple, Wikipedia knowledge graph-based methodology to build topic focused taxonomies. We utilize the Wikipedia graph and regard the taxonomy construction as a series of basic graph algorithms performed using topic-specific seed input nodes. Our case-studies demonstrate that the method performs well in general with respect to standard statistics derived from comparison with expert-curated manual taxonomies.

2 Methods and results

We construct the Wikipedia-based knowledge graph proposed and deployed by Aspert et al. [1] available at https://lts2.epfl.ch/Datasets/Wikipedia/. This graph is a directed multigraph with multiple nodes and edge types. Specifically, there are two classes of node which represent either Wikipedia articles or Wikipedia category articles (i.e. category pages). These in turn may be connected by two classes of directed edge which represent 'links_to' and/or 'belongs_to' relationships. A 'links_to' type edge connects two nodes if a hyperlink exists between the corresponding articles (the direction of the edge is straightforward), while a 'belongs_to' type edge represents a hyperlink between an article node or (sub)category node and a category node. For illustration, see Fig. 1. The graph is stored in a Neo4J graph database; for detailed description of the graph structure and other technicalities, see [1].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Fig. 1. Wikipedia graph structure. Blue (black) nodes: articles (as input). Green nodes: category pages. Black edges: hyperlinks connecting articles. Red edges: hyperlinks connecting articles or subcategories and parent categories.

2.1 Entity selection

The taxonomy generator is initialized with a collection of Wikipedia article type nodes $P = \{P^1, P^2, \dots\}$ and Wikipedia category type nodes $C = \{C^1, C^2, \dots\}$ which we process according to the following steps.

1. Construct a set $\mathscr{P} = \{P_1^1, P_2^1, \dots; P_1^2, P_2^2, \dots; \dots\}$ of all nodes which have a 'link_to' edge to one or more of the input pages, P.[4]
2. Start a depth-first traversal over each node $P_i^j \in \mathscr{P}$ for all 'belongs_to' type outgoing edges from P_i^j. At the first level this will result in the set $C_i^j = \{C_{i,1}^j, C_{i,2}^j, \dots\}$ of categories which the page P_i^j "belongs_to", at the second level the set of higher 'super'-categories of categories in C_i^j will be reached, and so on.)
   (a) If for a category node $C_{i,k}^j$, found during the traversal process $C_{i,k}^j \in C$ is satisfied, then add P_i^j to a "filtered" entity list $\mathscr{L}$;
   (b) Else, go to step 2, until all elements of $\mathscr{P}$ have been iterated over.

Note that in step 2. a stop criteria is required to restrict the maximum depth of the traversal process due to performance issues. In our experiments the criteria was set to a maximum depth level of four starting from the root node, provided that a category page in C had not already been reached.

2.2 Taxonomy creation

The next step is to classify each entity $e \in \mathscr{L}$ with a category and to provide a hierarchical category organization. For each e let C^e be the set of categories which e belongs to, that is, the neighborhood of e based on its outgoing 'belongs_to' type edges. Note that C^e is determined in step 2 of the entity extraction process. Let $\mathscr{C}$ be the set of all distinct

[4] After this step a fast filtering procedure can be applied by simply deleting any nodes from set $\mathscr{P}$ for which the node's corresponding Wikipedia page name either begins with a number (i.e. "2019_in_tennis") or contains the terms "by_year", "of_the_year", "List_of", or "_in_" (i.e. "Tennis_in_Hungary").

Taxonomy/category	TP/ Gold Players	TP/Gold Teams	TP/Gold All cat.	All Auto
American football	**90.65** (97/107)	**100** (32/32)	**74.59** (138/185)	4,068
Basketball	**89.15** (403/452)	**100** (30/30)	**89.67** (443/494)	5,526
Motorsport	**88.38** (784/88)	–	**86.12** (807/937)	5,862
Soccer	**79.8** (399/500)	**48.83** (294/602)	**62.96** (731/1161)	3,096
Tennis	**75.45** (206/273)	–	**67.17** (262/390)	2,077

Table 1. Coverage (ratio of true positives of automatically extracted entities and manually defined gold standard entities) results for several sports related taxonomies.

categories in $\bigcup_{e \in \mathscr{L}} C^e$. We define a bipartite graph over the disjoint node sets $\mathscr{L}$ and $\mathscr{C}$, where $e \in \mathscr{L}$ and $c \in \mathscr{C}$ are connected if e belongs to category $c \in \mathscr{C}$. Then, starting from c with the highest degree we greedily assign entities to categories step-by-step by removing the assigned entities and corresponding category in each step. Finally, to organize categories into a proper hierarchy one may use a pruning heuristics used e.g. in [4]

2.3 A case-study and evaluation

Domain-specific taxonomies are usually evaluated either by comparing them to manually-built (Gold Standard) taxonomies or by requesting feedback from experts in the field. One of our case-studies is targeted to build a taxonomy covering various sports.[5] Table 1 shows our experimental results regarding coverage (recall) values comparing the Gold Standard and automated taxonomy methods. It is noteworthy that the automated method finds many more relevant entities than the Gold Standard, however, for the purposes of this investigation this is a secondary concern to the primary aim of achieving a high recall compared to the Gold Standard. The high-precision reduction of irrelevant entities from the auto taxonomy (false positives) remains for future work.

References

1. Aspert, N., Miz, V., Ricaud, B., Vandergheynst, P.: A graph-structured dataset for Wikipedia research. In Proceedings of The 2019 World Wide Web Conference, pp. 1188-1193 (May 2019)
2. Hearst, M. A.: Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics, Vol. 2, pp. 539-545 (August 1992)
3. Ponzetto, S. P., Strube, M.:Taxonomy induction based on a collaboratively built knowledge repository. Artificial Intelligence, 175(9-10), 1737-1756 (2011)
4. Medelyan, O., Manion, S., Broekstra, J., Divoli, A., Huang, A. L., Witten, I. H.:. Constructing a focused taxonomy from a document collection. In Extended Semantic Web Conference, pp. 367-381 (May 2013)
5. Ponzetto, S. P., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In AAAI, Vol. 7, pp. 1440-1445 (July 2007)

[5]For example, for tennis the input Wikipedia page nodes were *Tennis, Association_of_Tennis_Professionals* and *Women's_Tennis_Association*, while the input category page node was *Tennis*.

The variability of network structures inferred from time series data

Mauro Faccin[1], Leto Peel[1], Alexandre Bovet[1], Benjamin Chiłm[1], Leonardo Gutierrez Gomez[1], Alexey Medvedev[1], Mridul Seth[1], and Jean-Charles Delvenne[1]

ICTEAM, Universit catholique de Louvain, Louvain-la-Neuve, Belgium

1 Introduction

Networks have become a popular representation for systems where distinct parts of the system are known to interact in a similar way. From the interaction between those parts we sometimes witness the emergence of a global behaviour which is not the sum of the atomic behaviours of each of those parts. We call those parts nodes and the interactions between them links. But for many complex systems, such as the neurological, financial and climate systems, we are not able to observe pairwise interactions directly and the researcher needs to resort to other means to infer the presence or absence of direct interaction between nodes. Once constructed, the network representation provides access to a range of network science tools and measures that can be used to characterize those systems.

There exist many ways to reconstruct a network. Most involve selecting a measure of similarity to compare the node signals and then selecting the most significant pairwise similarities to be represented as edges in the network. However, the reconstructed network and any subsequent analysis can be sensitive to the choices made in the reconstruction process, i.e., the measure of similarity, the method of selecting relevant edges and any parameter setting of the method. But often these choices are made arbitrarily and with little consideration of how sensitive results are to these choices. For instance, reconstruction methods are often tuned to achieve a arbitrarily specified density [1], obtaining a single connected component and/or to achieve a certain property, such as being small-world *enough*. Analogously in [2] the authors can set an hard threshold to recover the small-world property. In [3] the authors checks three values for the regularization parameter of the graphical LASSO [4] sparsification approach, to reconstruct a dense, medium and sparse graph. In [5] the authors show that in certain cases the community structure of the reconstructed networks does not widely depend on the threshold value selected. In this work we will explore the sensitivity of network statistics to some of these choices.

2 Results

We collected a number of time-series from different fields: MRI data for different tasks (neuroscience), the S&P100 price fluctuations (finance), local temperature changes (meteorology), tuberculosis quarterly reports (disease spreading). To compute a similarity

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

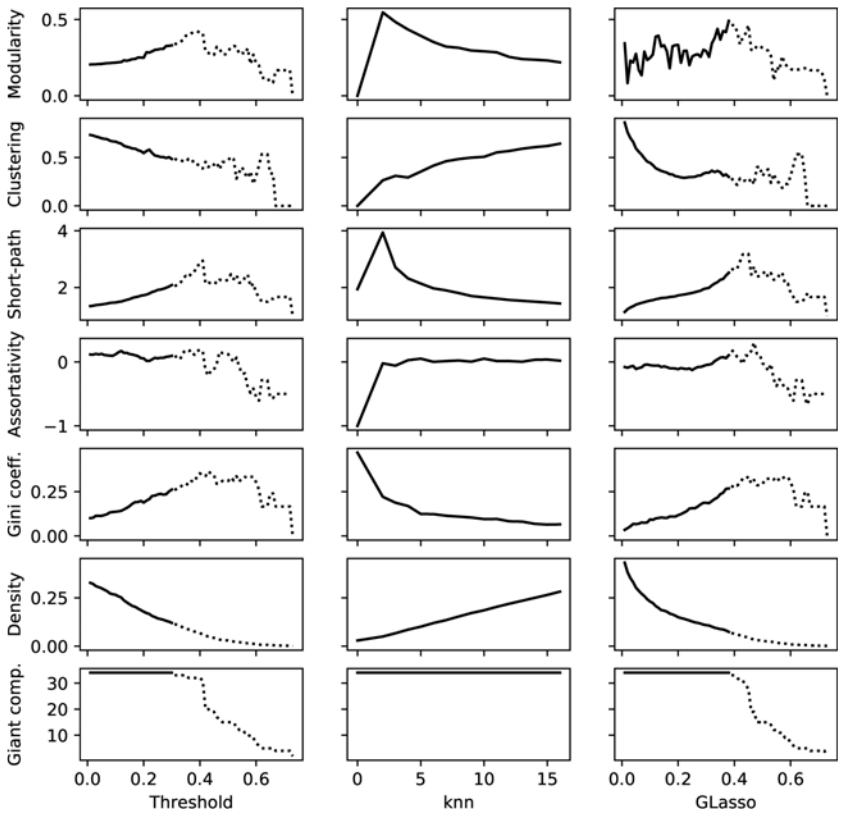


Fig. 1. Network statistics for different parameters of the sparsification method. The statistics fluctuate widely in the range of the sparsification parameter considered (the threshold value, the number of neighbours, the regularization parameter in graphical Lasso). Network statistics are computed on the giant component only, we use a solid line when the reconstructed network is composed by only one connected component. The network reconstruction is applied to the tuberculosis reports (quarterly) in health facilities in a sub–Saharan region.

measure between the time-series we select among several possibilities a popular approach that is both simple and pervasive: the Pearson's correlation.

To reconstruct the network from the similarity measure, and thus extract a sparse adjacency matrix, one has to choose which edges should be retained from all possible node–pairs. To this aim we apply a number of popular sparsification processes to the correlation matrix of the original time-series: (i) a fixed threshold; (ii) the k-nearest neighbours; (iii) the graphical Lasso [4]. In each case we consider undirected and unweighted networks without self-loops. In Figure 1 we visualize the network statistics in the case of tuberculosis reports in a sub-Saharan region, where each time-series represents a health zone in that region. For each reconstruction method we compute the modularity of the community structure found by the Louvain algorithm, the clustering coefficient, the average shortest-path length, the assortativity coefficient, the Gini [6] coefficient and the link density, all as function of the sparsification parameter. The network statistics of the reconstructed networks show wide fluctuations within the meaningful range of the sparsification parameters for all the reconstruction approaches. Despite the community structure may be robust [5], one can see the the modularity increase to almost double its value as the network gets more and more sparse and the weakest edges get removed. Similar analysis with similar results is performed for other datasets: the time-series of daily highest temperature reported in a number on US cities, the MRI scan of patients performing different tasks, the market values of the hundred leading US stocks in the S&P100 index.

Summary. From temporal activity data such as disease incidence reports, daily temperature and others, we perform sparsification of similarity matrix as found in many works in the literature in order to reconstruct the underlying network structure. The statistics of such reconstructed networks highly depend on the value of the sparsification parameter. Further work is necessary in this field to link analysis of temporal datasets to complex network techniques.

References

1. Lynall, M.E., Bassett, D.S., Kerwin, R., McKenna, P.J., Kitzbichler, M., Muller, U., Bullmore, E.: Functional connectivity and brain networks in schizophrenia. Journal of Neuroscience **30**(28) (2010) 9477–9487
2. Achard, S., Salvador, R., Whitcher, B., Suckling, J., Bullmore, E.: A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. Journal of Neuroscience **26**(1) (2006) 63–72
3. Rosa, M.J., Portugal, L., Hahn, T., Fallgatter, A.J., Garrido, M.I., Shawe-Taylor, J., Mourao-Miranda, J.: Sparse network-based models for patient classification using fmri. NeuroImage **105** (2015) 493 – 506
4. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3) (2008) 432–441
5. Yan, X., Jeub, L.G.S., Flammini, A., Radicchi, F., Fortunato, S.: Weight thresholding on complex networks. Phys. Rev. E **98** (Oct 2018) 042304
6. Gini, C.: Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1912)

Interpretability of model parameters in inference problems

Sergio Cobo-López[1], A. Godoy-Lorite[2] J. Duch[3], R. Guimerà[1,4], and M. Sales-Pardo[1]

[1] Departament d'Enginyeria Química, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain
[2] The Bartlett Centre for Advanced Spatial Analysis, University College London, First Floor 90 Tottenham Court Rd London London W1T 4TJ UK
[3] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain
[4] ICREA, 08010 Barcelona, Catalonia, Spain

sergio.cobo@urv.cat

1 Introduction

Making predictions about the unobserved behaviour of natural and social systems is a key element for scientific advancement. The massive generation of data in the society of information and the development of Machine Learning tools, provide an excellent opportunity to test our ability to make predictions in many different fields. In general, this is accomplished by exploiting patterns and regularities in those data with the aid of computer algorithms.

However, these algorithms are often difficult to interpret in their results and internal functioning. Our goal is precisely to develop models and their corresponding algorithms that make accurate predictions in a variety of contexts, but are easily interpretable. We consider two different datasets: one corresponding to a social experiment on strategic decision making, and another one consisting of the microbial content of human gut samples. The first dataset consists of a recent large-scale study of individuals playing a variety of 121 dyadic games in a controlled setting [1]. The second one is the result of a microbial analysis of stool samples of 883 patients.

In both cases, the data can be represented as a bipartite network with different types of links: there are two kinds of nodes in each set and interactions happen between nodes of different types; players perform a specific action when playing a game in the first example, and patient stool samples have a certain concentration of a specific microbe in the second example.

The common trait in both of these examples is that we expect that there are patterns in the way people play games and in the way microbes are distributed in a sample. We also expect that there is a finite set of patterns so that there exist groups of people that tend to perform the same type of action in the same games, and in the same way we

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

expect that there are groups of people who have similar concentration profiles of microbes in their gut. These common patterns can be thought of as being phenotypes that characterize individuals in different contexts. Our models thus work on the assumption that there exist groups of nodes in these networks. We capture them by exploiting the commonalities that exist in the behaviours of players or microbial profiles of patients (that is, in the connectivity patterns of nodes). Additionally, we can apply this mechanism to define groups of games and microbes. As a result, both problems considered can be tackled with the same group-based modeling approaches.

We implement two different models that belong to the Stochastic Block Model (SBM) family. The first one is a single phenotype model. Here, we consider that each person or microbe belongs to a single group and we look for the most plausible partition of groups. This partition is the one that maximizes the likelihood of the system and we use a simulated annealing to find it. In the second model, we allow nodes to belong to different groups simultaneously with different weights. That is, people and games or microbes have mixed memberships to different groups. As a consequence, groups are not any longer subsets of nodes. The membership profiles are encoded in what we call mixing vectors and we use an Expectation-Maximization algorithm to find the mixing vectors corresponding to the maximum likelihood. In both models, only three parameters are required: two vectors to track the membership of nodes to different groups or phenotypes (one vector for each species of nodes), and a matrix encoding the probability of connection between phenotypes. Therefore, it is very easy to understand why or why not the models succeed at making predictions by looking at these parameters. Furthermore, it is possible to analyze the dynamics of the system looking at the matrix of connections.

2 Results

Our results show that it is indeed possible to make reliable predictions on both problems highlighting the versatility and robustness of our inference approaches (Fig. 1 [a-b]). Not only that, but our results are easily interpretable in the single phenotype model as well as in the mixed phenotype one, which allows us to understand why predictions are accurate and to unveil key aspects of the underlying dynamics of the systems.
In the case of games, we conclude that (i) the perception of games by individuals is at odds with what should be expected from game theory; (ii) individuals tend not to follow single strategies, but rather mixtures of multiple strategies (Fig. 1 c).
In the case of microbes, we observe a well defined ecological order among groups of microbes and patients, characterized by the existence of increasing levels of specialization in their interactions (Fig. 1 d). This structure is called nestedness and it's very common in mutualistic networks. However, nestedness has not been observed in microbiome related systems before, at least to our knowledge.
As for the models used, we conclude that the mixed phenotype approach yields higher predictive accuracies and better interpretability regarding the interaction of different groups, albeit a higher number of parameters is generally required.

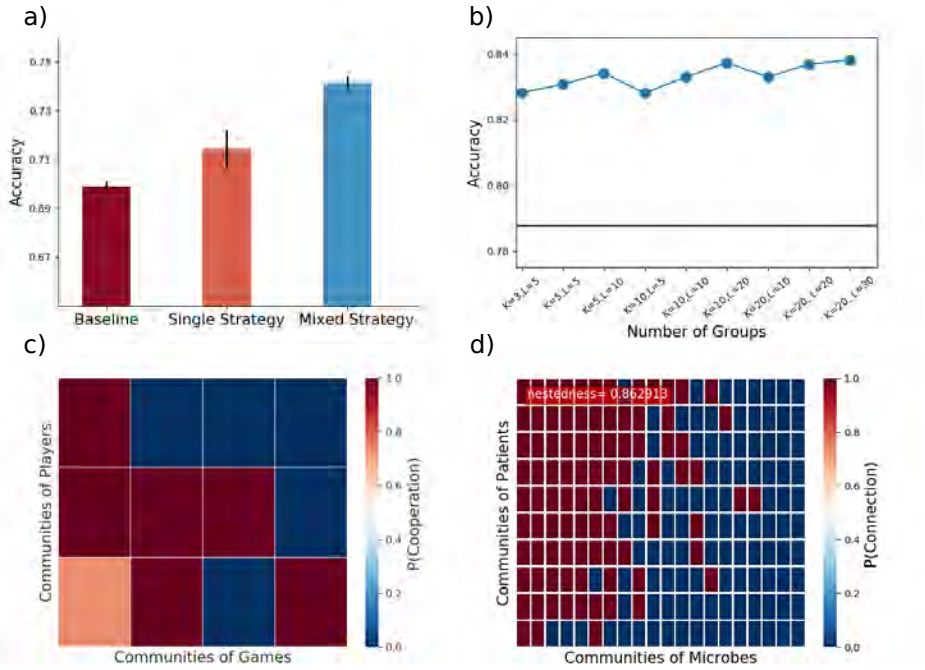


Fig. 1. (a) Predictive accuracy of the baseline model (red), the single phenotype (orange) and the mixed phenotype model (blue) for the game experiment. Each bin represents the average of a 5-fold cross-validation; error bars indicate the standard error of the mean. **(b)** Predictive accuracy of the mixed phenotype model for the human gut microbiome problem. Each circle represents the average of a 5-fold cross-validation for a given combination of K groups of patients and L groups of microbes. The black solid line indicates the baseline model. We observe a gradual and moderate increase in the predictive accuracy that saturates around $K = 10$, $L = 20$. **(c)** The relation of communities of patients and communities of games can be regarded as a set of latent strategies, indicating the behavioral pattern of groups of players towards groups of games. **(d)** The interaction between latent groups of patients and microbes shows a nested structure; specialist groups of patients contain a small subset of groups of microbes that are subset of those contained by more generalist groups of patients.

(1)

References

1. J. Poncela-Casasnovas, M. Gutiérrez-Roig, C. Gracia-Lázaro, J. Vicens, J. Gómez-Gardeñes, J. Perelló, Y. Moreno, J. Duch, and A. Sánchez, Sci. Adv. **2**, e1600451 (2016).

Rank Dynamics in Egocentric Social Networks

Sara Heydari[1], Gerardo Iñiguez[1,2,3], Jari Saramäki[1], and János Kertész[1,2]

[1] Department of Computer Science, Aalto University, Espoo, Finland,
sara.heydari@aalto.fi,
WWW home page: http://cs.aalto.fi/
[2] Department of Network and Data Science Central European University, H-1051 Budapest, Hungary
[3] IIMAS, Universidad Nacional Autonóma de México, 01000 Ciudad de México, Mexico

Egocentric networks (egonets) are fundamental units of human social networks. Thus, studying their structure helps to understand not only properties of people's personal networks, but also those of whole social network. An egonet consists of the ego as central node and his contacts (alters) connected to him via links. The structure of egonets and their evolution over time has been studied before. However, the majority of these studies investigate their evolution by comparing a few network snapshots [1], [2]. Here, to get more detailed insights on the dynamics of evolution of egonets, we study changes in egonets with high resolution, focusing on the dynamics of alter ranks.

The structure of egonets is often rather hierarchical, with few strong and many weak ties [3], [2], [4], [5]. The reason is that people divide their social attention heterogeneously among their contacts: a few alters get most of the communication time while the rest is distributed among many. Even though egonets are hierarchical, it seems that such hierarchies are not stationary in their composition and order. This was for instance reported in Ref. [2] where the observation timeline is divided into theree 6-months long consecutive periods and for each ego and each period the aggregated weighted egonet is constructed from phone call data. Following evolution of egonets, the study reports that from one period to the next, on average around 30% of top 20 alters are new. The mean turnover percentage is even higher for the complete egonets.

Large turnover in the membership of personal networks in consecutive time windows indicates that egonets are highly dynamic: social ties often decay or strengthen over time. Moreover, contacts on even stable and persistent ties are bursty [6], [7]. Here, instead of merely comparing snapshots of aggregated egonets, we closely follow the dynamics of egocentric networks with the highest possible resolution: event by event. To do so, we use a large mobile-phone call detail record (CDR) dataset which contains outgoing calls and text messages of egos during a 7-month period.

To form the timeline of an call egonet, we add call events one by one in chronological order. Therefore the link weights in the egonet (or the *scores* of alters) at time t are defined as the cumulative volume of communication since the beginning of the observation period, t_s. Fig. 1 shows the time evolution of the scores of alters of an example ego—one can clearly observe how some alters increase their rank through accumulating a lot of communication time.

Given the communication scores of alters of an ego at each time, t, we can rank the alters based on their scores. Then we can use a rank diversity measure to get an overview of the competition of alters in an egonet. Normalized rank diversity $d_r(t_s,t_e)$ of rank r is a measure of the number of unique elements which have been occupying that

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

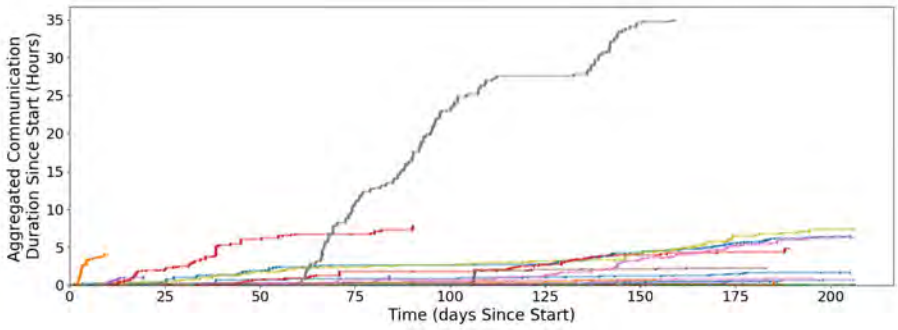


Fig. 1. The total time an ego has spent on phone with each of her alters from the beginning of data collection period till time t as a function of t. Here, we can observe how alters overtake others in terms of aggregated communication time. The heterogeneity of time allocation is also clearly visible, with one alter receiving a large fraction of communication.

rank between start time ,t_s, and end time ,t_e, and is defined as $d_r(t_s,t_e) := \frac{N_r(t_s,t_e)}{N(t_s,t_e)}$, where $N(t_s,t_e)$ is total number of alters that an ego has contacted during the whole period and N_r is number of alters who have occupied rank r in that period.

Plotting the rank diversities of egos as a function of (normalized) rank reveals that they generally have a parabolic shape (see fig. 2). This means that the highest and lowest ranks are visited by a few alters only, while most of the dynamics is going on in the middle ranks. This is different from the rank diversity curves observed in the context of usage-frequency of words in languages [8] or in competitive sports [10] where the bottom ranks have high diversity scores.

Our results point out that there might be a general mechanism that rules the rank dynamics of egocentric networks as they often have a parabolic rank-diversity curves. Nevertheless, we observed variation across the population which suggests that variations in the shapes of rank diversity curves might also be related to personality trait or demographic attributes of egos. For future research, we are interested in building models of communication which can explain the observed parabolic shape and also in searching our data for any correlation between attributes of egos and the specific shapes of their rank diversity curves.

References

1. Roberts, S. G., Dunbar, R. I., Pollet, T. V., & Kuppens, T.: Exploring variation in active network size: Constraints and ego characteristics. Social Networks, 31(2), 138-146 (2009)
2. Saramki, J., Leicht, E. A., Lpez, E., Roberts, S. G., Reed-Tsochas, F., & Dunbar, R. I.: Persistence of social signatures in human communication. Proceedings of the National Academy of Sciences, 111(3), 942-947 (2014)
3. Onnela, J. P., Saramki, J., Hyvnen, J., Szab, G., Lazer, D., Kaski, K., ... & Barabsi, A. L.: Structure and tie strengths in mobile communication networks. Proceedings of the national academy of sciences, 104(18), 7332-7336 (2007)

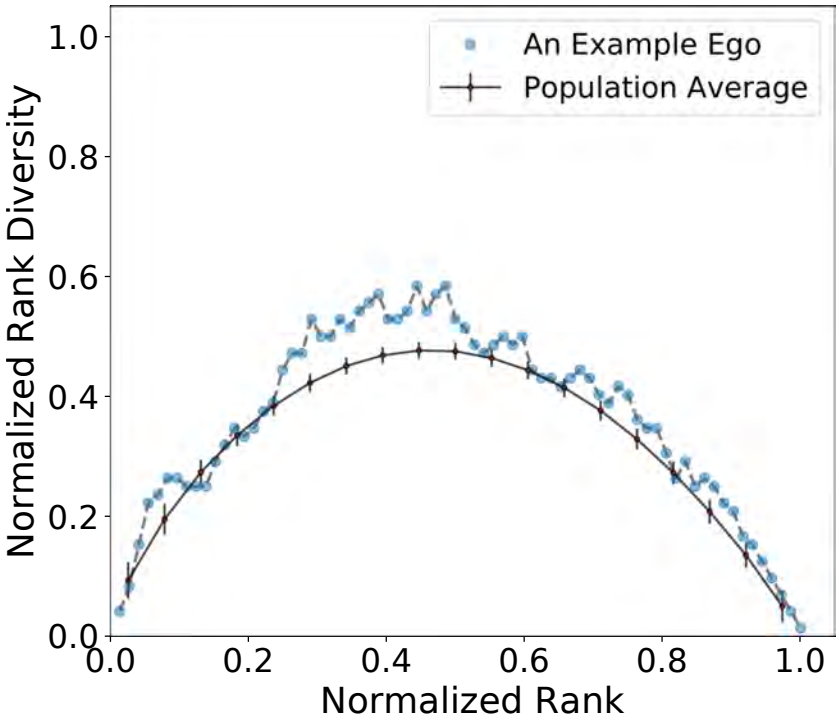


Fig. 2. In blue, you can see normalized-rank diversity curve of an example ego as a function of *normalized rank*. If the rank set in an egonet is $\{1,2,8\ldots N\}$, then $\{\frac{1}{N},\frac{2}{N},8\ldots 1\}$ is its corresponding normalized rank set. Defining normalized rank makes ranks across egonets of different size comparable and enables us to calculate population- average. In red, you can see population average of normalized rank-diversity (sample size 8000) as a function of normalized rank which has an almost perfect parabolic shape. The errorbars are equal to one standard deviation.

4. Miritello, G., Moro, E., Lara, R., Martnez-Lpez, R., Belchamber, J., Roberts, S. G., & Dunbar, R. I.: Time as a limited resource: Communication strategy in mobile phone networks. Social Networks, 35(1), 89-95 (2013)
5. Heydari, S., Roberts, S. G., Dunbar, R. I., & Saramki, J.: Multichannel social signatures and persistent features of ego networks. Applied network science, 3(1), 8 (2018)
6. Karsai, M., Kivel, M., Pan, R. K., Kaski, K., Kertsz, J., Barabsi, A. L., & Saramki, J.: Small but slow world: How network topology and burstiness slow down spreading. Physical Review E, 83(2), 025102 (2011)
7. Navarro, H., Miritello, G., Canales, A., & Moro, E.: Temporal patterns behind the strength of persistent ties. EPJ Data Science, 6(1), 31 (2017)
8. Morales, J. A., Snchez, S., Flores, J., Pineda, C., Gershenson, C., Cocho, G., ... & Iiguez, G.: Universal temporal features of rankings in competitive sports and games. arXiv preprint arXiv:1606.04153 (2016)
9. Penrose, M.: Random Geometric Graphs. Oxford studies in probability, O.U.P. (2003)
10. Cocho, G., Flores, J., Gershenson, C., Pineda, C., & Snchez, S.: Rank diversity of languages: Generic behavior in computational linguistics. PLoS One, 10(4), e0121898 (2015)

Latin Text Authorship using Dynamical Measurements of Complex Networks

Clara Gracio[1], Luis Garcia Zapata[2], Ligia Ferreira[3], Irene Rodrigues[3], Claudia Teixeira[4], and Armando S. Martins[4]

[1] University of Evora, ECT, Mathematics Department and CIMA, Portugal
[2] Departamento de Matematicas, Universidad de Extremadura, Espanha
[3] ECT, Departamento de Informatica e LISP, Universidade de Evora, Portugal
[4] ECS Departamento de Linguas e Literatura, Universidade de Evora, Portugal

1 Introduction

Recognition of authorship and literary style of a writer has been a frequent subject of investigation. Today, a new approach to representing and modeling complex systems has gained strength and has proven powerful: complex networks. They have modeled many real systems from the internet to the human body, including texts. The most used textual network for recognition of authorship is the co-occurrence of words. In this paper, we study a Latin text network to obtain authorship verification [2, 1]. The Latin texts are from Historia Augusta, a collection of biographies of Roman emperors extending from Hadrian (117-138) to Carus (282-83) and his son Carinus (283-285) and Numerian (283 (284), and ours. Traditionally, the work is attributed to six different authors (collectively known as the Scriptores Historiae Augustae). The true authorship of the work, its actual date, its reliability, and its purpose, have long been matters for controversy amongst historians and scholars, ever since Hermann Dessau in 1889 rejected both the traditional date and authorship. The objective of this work is to verify the hypothesis that the attribution of authorship of the texts is correct (six authors) or that all texts were written by one author.

We construct a method that combines traditional measurements and complex network measurements. Traditional methods that use frequency of words, characters and character bigrams like Stylo with R and k-means. K-means is a partial algorithm that is the most commonly used because is fast and produces good results, see fig 1 for Latin Ha.

The proposed method for authorship attribution is based on the evolution of the topology of networks, i.e. we exploit the network dynamics[4]. Therefore, unlike previous approaches we do not construct one single network from the whole book. Instead, a book is divided into shorter pieces of text [3] comprising the same number of words. Then, a co-occurrence network is constructed for each part, which generates a series of independent network for each book.

In fig 2 we present the graph for the following words co-occurrence: duobus, duobus $->$ liberis, liberis$->$ quos, quos$->$SeptimiusSeverus, SeptimiusSeverus $->$reliquit , reliquit $->$Getam,Getam$->$ et , et$->$Bassianum, Bassianum $->$quorum, quorum $->$unum, unum$->$ Antoninum, Antoninum$->$ exercitus, exercitus $->$alterum, alterum$->$ pater, Pater$->$ dixit, dixit$->$Geta, Geta$->$hostis, Hostis$->$est, est$->$ iudicatus, iudicatus$->$ Bassianus , Bassianus $->$autem, autem$->$ obtinuit, obtinuit$->$ imperium

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

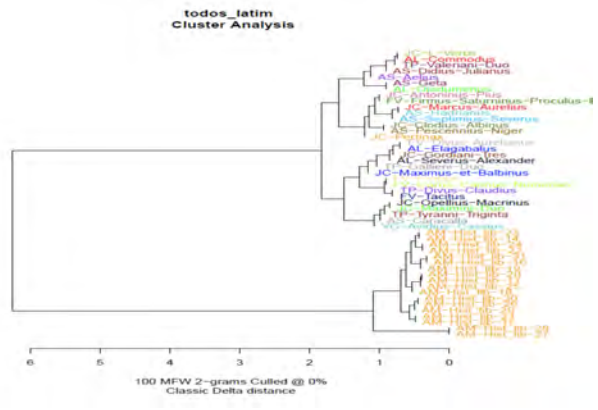


Fig. 1. stylo analysis

Fig. 2. Co-occurrence graph

2 Results

In our work we used two different preprocessing steps before transforming texts into networks, first we build our graphs by removing all punctuation marks from the text and second we build our graphs by removing all punctuation marks from the text and all stopwords . Then, a co-occurrence network is constructed for each part, which generates a series of independent networks for each book. Each partition is described by the following topological network measurements[5, 6]: clustering coefficient, network diameter, network radius, number of cliques, betweenness centrality, shortest path length, degree centrality, total number of nodes, total number of edges, the graph assortativity and (for the first time in a co-occurrence network) the second-smallest eigenvalue (counting multiple eigenvalues separately) of the Laplacian matrix, the Fiedler value. In fig 3 we considered a book divided into 182 parts with 500 words each, without stopwords.

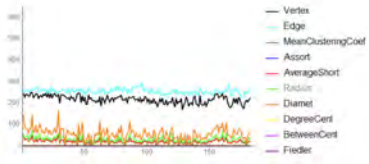


Fig. 3. Series of book Graph parametres

To create a mathematical model credible based on complex network parameters, it was necessary to consider other texts whose authorship is known and there is no doubt. In order to compare with completely different dates and styles, we also considered some Portuguese authors: Jos

Saramago, Mia Couto, Antnio Lobo Antunes. We introduce a classifier that uses the previously calculated network parameters. K-nearest neighbors (KNN), that infers the class of an instance by a voting process over the nearest neighbors in the training dataset. Our experiments on portuguese books, 3 authors with 3 books each, show that the co-occurrence graphs that we build for each book, 20 graphs for text peaces of 200 and 500 words, can be well classified with the algorithm of K-nearst neighbors. Our tests enable us to conclude that this method is very good in detecting authorship for the portuguese authors. We train the 8 books(2 ALobo, 3 Mia, 3 Sara) and test with the book of one author (ALobo), each book is represented by 20 graphs characteristics, 500 Words (with stop words).

knn.1	ALobo	Mia	Sara
ALobo	20	0	0
Mia	0	0	0
Sara	0	0	0

knn.1	Alobo	Mia	Sara
Alobo	0	0	3
Mia	0	0	1
Sara	0	0	16

knn.1	Alobo	Mia	Sara
Alobo	0	2	0
Mia	0	13	0
Sara	0	5	0

Our experiments, see table below, with Historia Augusta, 3 authors (5 texts, 4 texts and 9 texts), for Historia Augusta authors it is not conclusive.

knn.1	AS	JC	TP
AS	8	0	0
JC	6	0	0
TP	6	0	0

knn.1	AS	JC	TP
AS	0	0	4
JC	0	0	14
TP	0	0	0

For some authors it is possible to recognize the authorship But some tests fail this task

We can observe that we cannot say that there are six different authors, maybe results confirm that the study in [2] is correct.

Keywords: Latin text Authorship, complex networks, topological parameters, co-ocurrence networks, spectral clustering.

Acknowledgements This work has been partially supported by (CIMA) through the grant UID/MAT/04674/2013, by (LISP) through the Grant UID/CEC/4668/2016, both research centers are supported by FCT (Fundao para a Ciłncia e a Tecnologia, Portugal) and, also, by Dep. de Matemticas, Escuela Politcnica de Cceres, de la Universidad de Extremadura, Spain.

References

1. TEIXEIRA, Cludia; RODRIGUES, Irene. Deciphering Latin sentences using traditional linguistic resources. Digital Scholarship in the Humanities, 2018.
2. Stover, J. A., Kestemont, M. (2016). THE AUTHORSHIP OF THE HISTORIA AUGUSTA: TWO NEW COMPUTATIONAL STUDIES. Bulletin of the Institute of Classical Studies, 59(2), 140-157.
3. Tohalino, J. V., Amancio, D. R. (2017, October). Extractive Multidocument Summarization Using Dynamical Measurements of Complex Networks. In 2017 Brazilian Conference on Intelligent Systems (BRACIS) (pp. 366-371). IEEE.
4. Newman, M., Barabasi, A.L., and Watts, D.J. The structure and dynamics of networks. Princeton University Press, 2011.
5. Kannan, R., Vempala, S., Vetta, A. (2004). On Clusterings: Good. Bad and Spectral. Journal of the ACM, v.51, pp. 497-515.
6. J. Leonel Rocha, Sara Fernandes, Clara Grcio and Acilina Canec0, Spectral and Dynamical Invariants in a Complete Clustered Network, Appl. Math. Inf. Sci. 9, No. 6, 1-10 (2015).

A Network model of the Chemical Space provides similarity structure to the system of chemical elements

Eugenio Llanos[1,2,3], Wilmer Leal[1,2] Andrés Bernal[2,4], Guillermo Restrepo[2], Jürgen Jost[2], and Peter F. Stadler[1,2,5]

[1] Bioinformatics Group, Department of Computer Science, Universität Leipzig,Härtelstraße 16-18, 04107 Leipzig, Germany
ellanos@sciocorp.org,
[2] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[3] Corporacióon SCIO, Calle 57b 50-50 bloque d22 of. 412, 111321 Bogota, Colombia
[4] Departamento de Ciencias Básicas, Universidad Jorge Tadeo Lozano, Carrera 4 22-61, Bogota, Colombia
[5] The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico 87501

1 Introduction

The collection of every species reported up to date constitutes the so-called Chemical Space (CS). This space currently comprises well over 30 million substances and is growing exponentially[2]. In order to characterize this ever-growing space, chemists seek for similarity of substances on the CS based on the way they combine[3]. Mendeleev's work on chemical elements was based upon his knowledge of the CS by 1869 is perhaps the most famous example of how the CS determines similarity relations [4]. From a contemporary point of view, Network Theory serves as a natural framework to identify these kind of relational patterns in the CS [5]. Nowadays, databases such as Reaxys have grown to a point where they can be taken as proxies for the whole CS, opening the possibility to analyze it from a data driven perspective.

In this work we propose to study the similarity of chemical elements according to the compounds they form. From each compound, we deleted each element to obtain a formula that is connected to the deleted element, v.g. $S_{1/2}O_{4/2}$, $Na_{2/1}O_{4/1}$ and $Na_{2/4}S_{1/4}$ are formulae coming from Na_2SO_4 (Sodium sulfate) where Na, S and O, have been deleted respectively. This form a bipartite graph formed by elements and those formulae where they have been deleted, We build our network using 26,206,663 compounds recorded on Reaxys up to 2015. Similarity among chemical elements is constructed analogously to Social Network Analysis, where actors are declared similar whenever they are connected to the same set of other actors. The more formulae elements share, the more similar they are. We introduce a new notion of in-betweenness of elements acting as *mediators* on similarity relations of others. We analyze the structural features of this network and how they are affected by node removal. We show that the network is both highly dense and redundant. Even though it is heavily centralized, similarity relations are widely spread across a wide range of formulae, which grants the network extraordinary structure resiliency, even against directed attack. We discuss some implications of these results for chemistry.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

- *The network is heavily centralized*: chemical reactivity of elements is far from uniform, as the degree distribution of elements exhibits three different regions, see Figure 1(b). The first one is composed by a few elements that concentrate the vast majority of relations (the first is H, which accounts for 95.9% of formulae, followed by C 95%.0). The second one is composed by the bulk of elements, which connect to 10,000-100,000 formulae. The third region corresponds to elements that have a very low number of molecular formulae.
- *Formulae with degree one are mostly connected to central elements*: formulae of degree one correspond to compounds that are unique to one element (*singularities*). Eight elements concentrate most singularities (90%) evoke both the singularity principle of the periodic chart and the distinction between organic and inorganic chemistry. In general, the number of singularities scales semi-linearly with element degree (power law with exponent 1.3, see Figure 1(a)). This result shows that elements tend to be unique as long as more compounds of them are obtained, independently of their identity. The more compounds one element has, the less similar to others it becomes.
- *Similarity does not partition the space into clear-cut classes of elements*: since formulae generate similarity relations among the elements that are connected to them, the degree of one formula corresponds to the number of elements it makes similar. The smoothness of this degree distribution (Figure 1(a)) shows that elements cannot be divided into clear-cut classes, since otherwise such classes would produce local maxima corresponding to the sizes of these classes. This result has an interesting chemical implication, as it challenges the usual view of elements as separated *families*.
- *Element in-betweenness depends on its degree*: elements work as mediators of similarity relations through the formulae they constitute. Such mediation scales almost linearly with the degree of the element (see Figure 1(c)). This is a very interesting feature, since it shows that similarity relations are not concentrated on certain kind of compounds or manifested by specific elements working as mediators, but are evident on the entire CS.
- *Similarity relations are highly resilient to directed attack*: since the network is highly centralized, deleting random elements should not have a major effect on the network topology. We instead deleted sequentially elements from the one with highest degree down to 12 elements and those formulae on which they take part. Deleting central elements has impact on the degree of the elements and the distribution goes down on absolute frequency. Notwithstanding, almost all elements are affected in the same way and the shape of the curve is conserved (see different data series on Figure 1(b)). The same happens on the degree of formulae, which is shifted towards the left, but the shape remains (Figure 1(a)).
- *Strong and weak similarity relations are the less variant*: since our network is of an epistemic nature, vulnerability can be related to the viability of extracting knowledge with limited information. To test how variant are the similarity relations against removal of molecular formulae, we calculated the variance of the rank of pairwise element similarity (number of length 2 paths between the correspond-

ing nodes) when keeping only similarities mediated by each element. Surprisingly, strong and weak similarities have the lowest variance (see Figure 1(d)), showing that similarities are by no means random but they form a strong structure that stands across the entire CS, revealing a fundamental nature of these similarity patterns.

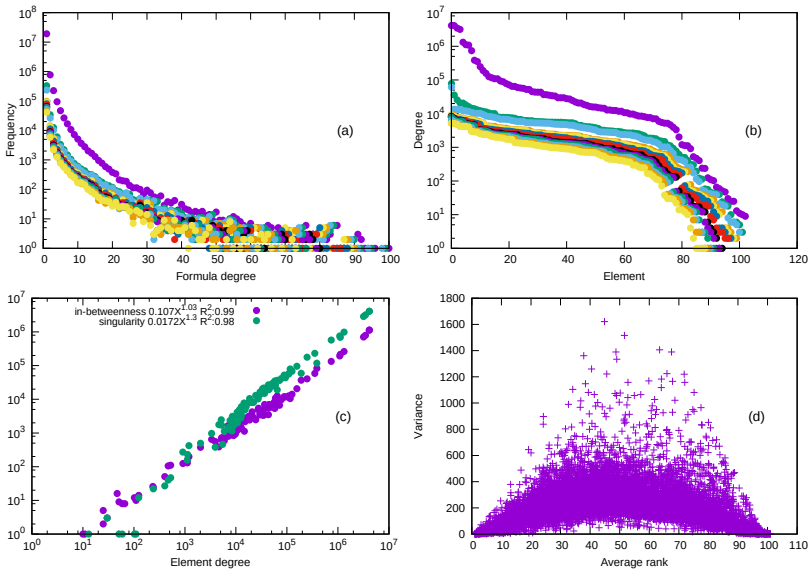


Fig. 1. (a) Distribution of formula degree. Different colors of points correspond to series where different central elements has been removed. (b) Degrees of elements. (c) Singularities and in-betweennes vs formula degree. (d) Variance of pairwaise rank position vs average rank position. Low variance is found on low average rank positions (similar elements) and high average rank positions (dissimilar elements).

References

1. Schummer, J.: Scientometric studies on chemistry II: Aims and methods of producing new chemical substances. Scientometrics 39 (1), 125–140 (1997)
2. Llanos, Eugenio J.; Leal, Wilmer; Luu, Duc H.; Jost, Juergen; Stadler, Peter F.; Restrepo, Guillermo: Exploration of the chemical space and its three historical regimes. Proceedings of the National Academy of Sciences of the United States of America 116 (26), 12660-12665 (2019) https://doi.org/10.1073/pnas.1816039116
3. Schummer, J.: The chemical core of chemistry I: a conceptual approach. HYLE 4, 129-162 (1998)
4. Leal, Wilmer; Llanos, Eugenio J.; Stadler, Peter F.; Jost, Juergen; Restrepo, Guillermo: The Chemical Space from Which the Periodic System Arose. ChemRxiv. Preprint (2019)
5. Leal, Wilmer; Restrepo, Guillermo; Bernal, Andrés. A network study of chemical elements: from binary compounds to chemical trends. MATCH communications in mathematical and in computer chemistry 68, 417442 (2012)

An empirical study of the relation between the overlapping nodes and hubs in networks with modular structure

Z. Ghalmane[1,3], M. El Hassouni[1], C. Cherifi[2], and H. Cherifi[3]

[1] LRIT, Mohammed V University, Rabat, Morocco,
[2] DISP Lab, University of Lyon 2, Lyon, France
[3] LE2I, University of Burgundy, Dijon, France
zakaria.ghalmane@gmail.com

1 Introduction

Community structure is one of the most organizing properties of real-world networks. There is no universal definition of the community structure. Yet, it is usually described as groups of densely connected nodes with loose connections to nodes from different modules. Two types of community structure can be found in the literature depending on the nature of nodes. The majority of networks exhibit an overlapping community structure where nodes may belong to multiple communities [1]. In real-world networks, hubs are nodes with a number of links that greatly exceeds the average. They play a major role in terms of information dissemination. In previous work, M. Kumar *et al.* [2] used the OverlapNeighborhood strategy to highlight the most connected nodes in the network. This strategy selects randomly the immediate neighbors of the overlapping nodes for immunization. It is based on the idea that overlapping nodes are part of several communities. Therefore, there is a high chance that nodes with high degrees to be a neighbor of overlapping nodes. This strategy targets hubs by requiring only information at the level of overlapping nodes and without the knowledge of the global structure of the network. It supposes that overlapping nodes are neighbors of the highly connected nodes. Our aim in this paper is to confirm this assumption. Few studies have been interested in the topological analysis of overlapping nodes [3]. In this work, we try to measure the proportion of hubs located in the immediate neighborhood of the overlapping nodes. This is in an attempt to understand the relation between the overlapping nodes and the hubs. Experiments performed on empirical networks with overlapping community structure show that the hubs represent a large proportion of the immediate neighbors of overlapping nodes.

2 Materials and Methods

Methods. To study the relationship between the hubs and the overlapping nodes, two empirical approaches are adopted. First, we compute the proportion of the hubs in the list of neighbors of the overlapping nodes according to the algorithm 1. If this proportion is greater than 50%, we consider that a high portion of the hubs are neighbors to the

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

overlapping nodes. In the second approach, the Spearman correlation coefficient is used to quantify the relationship between both hubs and the neighbors of the overlapping nodes. We consider X and Y as the rank vectors of nodes belonging respectively to the ordered list of hubs and the ordered list of neighbors of the overlapping nodes. The ranks are computed according to the degree of nodes. The Spearman correlation coefficient ρ between the two vectors X and Y is defined as:

$$\rho(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

The following ranges are considered to measure the strength of the relation between the hubs list and the neighbors list: It is a moderate correlation if the coefficient ranges between 0.5 and 0.7, a high correlation if the coefficient ranges between 0.7 and 0.9, and a very high correlation if it ranges between 0.9 and 1. Therefore, our hypothesis is considered verified if both measures have a quite high values. In this case, we confirm that the hubs are neighbors to the overlapping nodes.

Algorithm 1: Computation of the proportion of hubs in the neighborhood of the overlapping nodes

Input : Graph $G(V,E)$, Number of nodes $n \leftarrow |V|$, List of the overlapping nodes L_o
Output: Proportion of hubs in the neighborhood of the overlapping nodes p

Create and initialize the list of neighbors of the overlapping nodes L_{on}
for *each* $v \in V$ **do**
  Add all the neighbors of the node v to the list L_{on}
end
Initialize k the size of neighborhood of the overlapping nodes : $k \leftarrow size(L_{on})$
Sort all the nodes of the network in decreasing order according to their degree
Create and initialize the list of hubs L_h
Add the top k nodes of the network to the list L_h
$p \longleftarrow 0$
for *each* $v \in L_h$ **do**
  if $v \in L_{on}$ **then**
    $p \longleftarrow p+1$
  end
end
$p \longleftarrow p/n$
Return p

Empirical datasets. A set of four real-world networks is considered to perform a series of experiments. The selected networks are from various origin (social, co-appearance, collaboration and e-commerce networks) and different sizes to cover a wide range of situations. The overlapping community structure of the networks is discovered using the the Speaker-Listener Label Propagation Algorithm SLPA. This algorithm is chosen because of its good compromise between the effectiveness and the complexity when used in many different types of networks [1].

-Zachary's karate club: is a social network of friendships between 34 members of a karate club at a US university in the 1970s.
- Les Miserables: is a coappearance network of characters in the novel Les Miserables.
- ca-GRQC: is a collaboration network which has been collected from the e-print arXiv. It covers scientific collaborations between authors of papers submitted to the General Relativity and Quantum Cosmology category.
-Amazon co-purchasing network: This network is collected from Amazon web site. If a product i is frequently co-purchased with product j, the graph contains an undirected edge from i to j.

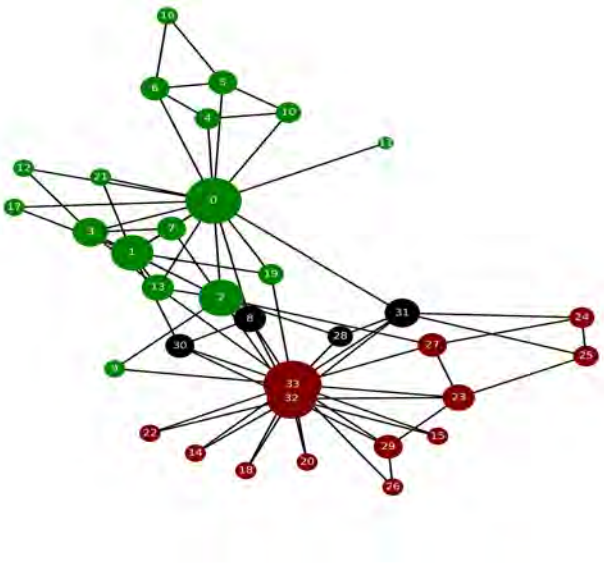


Fig. 1. Karate club network. Nodes with the same color belong to the same community. Nodes in black represent the overlapping nodes.

3 Results

In this section, we measure the proportion of hubs in the neighbors of the overlapping nodes. The correlation between the list of hubs and the list of the neighbors of the overlapping nodes is also computed. The experimental results are performed on four networks of various nature. The results are reported on Table 1.

Table 1. Statistics and properties of the real-world networks. N is the total numbers of nodes. E is the number of edges. on is the number of the overlapping nodes. N_c is the number of communities. p represents the proportion of the hubs in the neighborhoods of the overlapping nodes. c is the correlation between the list of hubs and the list of the neighbors of overlapping nodes.

Network	N	E	on	N_c	$p(\%)$	$c(\%)$
Zachary's karate club	34	78	4	2	77.6	99
Les Miserables	77	254	5	5	81.06	97.2
ca-GrQc	4158	13428	1138	667	65,04	94.6
Amazon	334863	925872	81215	75149	68,31	96

Figure 1 illustrate the Karate club network. This network has two communities and four overlapping nodes (colored in black). It is noticed from this figure that all the overlapping nodes are connected to the hubs belonging to both communities. We note that the size of nodes is assigned according to their degree. Thus, a high fraction of hubs takes part of the neighborhood of the overlapping nodes in the Karate club network. Moreover, the proportion of the hubs among the overlapping neighborhood is always greater than 50% for all the other networks. However, it shows the highest value in networks with small sizes. Additionally, for all the tested networks, there is a very high correlation between the list of hubs and the list of neighbors of the overlapping nodes. Therefore, a high proportion of hubs are neighbors to the overlapping nodes for the four tested networks.

To summarize, a set of experiments are performed on empirical networks to characterize the relation between the hubs and the overlapping nodes. Results show that a high proportion of the hubs take part of the neighborhoods of the overlapping nodes. This confirm the assumption that overlapping nodes are neighbors with the highly connected nodes of the network. Further analytic work should be performed to define the exact nature of this relationship.

References

1. Jebabli, M., Cherifi, H., Cherifi, C., & Hamouda, A. (2018). Community detection algorithm evaluation with ground-truth data. Physica A: Statistical Mechanics and its Applications, 492, 651-706.
2. Kumar, M., Singh, A., & Cherifi, H. (2018, April). An efficient immunization strategy using overlapping nodes and its neighborhoods. In Companion of the The Web Conference 2018 on The Web Conference 2018 (pp. 1269-1275). International World Wide Web Conferences Steering Committee.
3. Yang, J., & Leskovec, J. (2012). Structure and overlaps of communities in networks. arXiv preprint arXiv:1205.6228.

Part XI

Network Geometry

Connectivity of 1-dimensional Soft Random Geometric Graphs

Michael Wilsher*, Carl P Dettmann, and Ayalvadi Ganesh

School of Mathematics, University of Bristol, Bristol, BS81TW, UK
michael.wilsher@bristol.ac.uk
**presenting author*

1 Introduction

The original (hard) random geometric graph (RGG), as proposed in [1], was created by generating a Poisson point process (PPP), Φ, on some space, $\mathscr{V} \subset \mathbb{R}^d$, and then placing an edge between any two nodes a distance at most r_c apart. This idea was formulated as a way of adding a spatial element to the Erdős-Rényi random graph. These RGGs have been well studied and used in many different applications for which this spatial element is important. These include modelling disease spread in social networks, climate, infrastructure, and neuronal networks. A more general review on spatial networks is given in [2]. The original idea, however, was to use these graphs to model ad-hoc wireless communication systems and this model has been widely used in this area (see [3, 4] for example). The idea in an ad-hoc wireless communication system is that devices communicate with each other directly, rather than via a central router, therefore allowing for increased mobility and scalability, and removes the single point of failure from the network. The nodes in our RGG represent users or devices and an edge between any pair of nodes indicates that communication is possible between these two users/devices.

In the one-dimensional RGG model nodes are distributed on the line according to some random process and edges are placed between nodes a distance at most r_c apart. One main application has been to vehicular ad-hoc networks (VANETs) [5]. Here the line represents the road, the nodes represent the vehicles, and an edge between two nodes indicates that two vehicles are able to communicate with each other.

The model of RGGs has also been extended to look at a non-binary version of connectivity where edges are now created with a probability that is a function of the distance between the points. These have many different names in the literature but we will call them "soft" RGGs (SRGG) [6, 7]. These have also been called "random connection models" and Waxman graphs. This new model adds an extra layer of randomness to these graphs and also creates a more realistic model of physical networks where connectivity is not as simple as being within a certain distance. This allows us to model fading into our model in the form of a probabilistic connection function. In this new model, the point process is now generated in the same way, however now an edge is placed between nodes in our network with probability $H(r)$ where r is the mutual distance between the nodes. The general form of these connection functions is that they are smoothly decreasing from 1 to 0 as r is increasing from 0 to ∞. Since the main application of this work is to ad-hoc communication networks and, in particular, to vehicular communication systems, we have looked at connection functions of the form

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

$$H(r) = \exp\left(-\mu r^{\eta}\right). \tag{1}$$

Here η is the path loss exponent and μ is a constant that determines the length scale. η is based on the type of environment in which our network is placed. Typically for a line of sight communication, η takes a value of 2, for more dense urban environments η takes values of 3 or 4 (going as high as 6 or 7 for a city such as Manhatten). A value of $\eta = 1$ would represent an environment in which there are low levels of reflection such as vehicles in a tunnel. For an overview of possible connection functions for different wireless communication systems, see [7].

A widely asked question when looking at these graphs is what is the probability of having a fully connected network. In other words, what is the probability of having a path (either single- or multi-hop) from every node in the graph to every other node. When looking at SRGGs in dimension $d \geq 2$, a powerful result is given in [6] stating that in the limit of the number of nodes going to infinity and under certain conditions on the rate of fading of the connection function, the only obstruction to full connectivity is isolated nodes. It was also shown that the number of isolated nodes in this graph can be well approximated by a Poisson distribution. Therefore, to find the probability of full connectivity, one simply needs to find the probability of having no isolated nodes in the network, which is given by

$$\mathbb{P}(N_{\text{iso}} = 0) = e^{-\mathbb{E}[N_{\text{iso}}]} \tag{2}$$

where N_{iso} is the number of isolated nodes in our SRGG. This result, however, doesn't hold in the case of $d = 1$. This is due to the fact that in dimension one, the graphs can "split" into separate large clusters, a phenomenon not witnessed in the higher dimensional version. Therefore, a different method for calculating the probability of full connectivity is required.

2 Model and Results

For our work, the model is as follows. Generate a Poisson point process of rate = 1 on a line segment of length L. Each pair of nodes is then connected independently with probability $H(r)$ as defined in Eqn. (1). We concentrate on $\eta = 1, 2$ which are respectively called the Waxman and Rayleigh connectivity functions. In our work, we are firstly looking at a large but finite L and varying μ in our connection function. The idea of looking at a large but finite L is that this will best represent a vehicular network on a highway. Secondly, we look at a large L limit in which μ scales with L. Within these two different regimes, we wish to find the probability of having a fully connected network. For this to happen, we need to have none of the following three events: isolated nodes, uncrossed gaps, and splits. These are illustrated in Figure 1.

In the work so far, we have seen that the two dominating factors in connectivity are the isolated nodes and uncrossed gaps. We have calculated the expected number of isolated nodes in our network and shown via. simulations that the number of isolated nodes follows a Poisson distribution. This means that we have a very good approximation for the probability of having no isolated nodes in the network. We have also shown via. simulations that this same behaviour occurs when calculating the probability of having

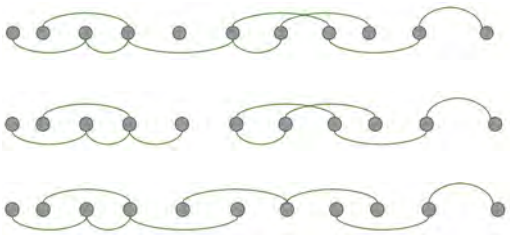


Fig. 1. The three different ways in which a disconnection can occur in our network are isolated nodes (upper), uncrossed gaps (middle), and splits (lower).

no uncrossed gaps in the network. Since these are the main two contributing factors to not having a fully connected network, the probability of full connectivity can be approximated by one minus the probability of having neither of them occur. This leads to the following equation for the probability of full connectivity in our network model.

$$P_{\text{fc}} \approx \exp(-\mathbb{E}[N_0] - \mathbb{E}[N_{\text{ucg}}]) \tag{3}$$

where $\mathbb{E}[N_0]$ and $\mathbb{E}[N_{\text{ucg}}]$ are the expected number of isolated nodes and uncrossed gaps respectively.

Summary. In summary, one dimensional soft random geometric graphs are surprisingly more complicated to analyse than their higher dimensional analogues. Furthermore, an understanding of one dimensional connectivity is needed for the development of effective vehicular peer to peer communications networks.

References

1. Gilbert, E.N.: Random plane networks. J. Soc. Ind. Appl. Math. 9(4), 533–543 (1961)
2. Barthlemy, M.: Spatial networks. Phys. Rep. 499(13), 1 – 101 (2011)
3. Haenggi M, Andrews JG, Baccelli F, Dousse O, Franceschetti M.: Stochastic geometry and random graphs for the analysis and design of wireless networks. IEEE Journal on Selected Areas in Communications. 27(7):1029–1046 (2009).
4. Dettmann CP, Georgiou O, Pratt P.: Spatial networks with wireless applications. Comptes Rendus Physique. 19(4):187-204 (2018).
5. Knight G, Kartun-Giles AP, Georgiou O, Dettmann CP.: Counting geodesic paths in 1-D vanets. IEEE Wireless Communications Letters. 6(1):110–113 (2016).
6. Penrose MD. Connectivity of soft random geometric graphs. The Annals of Applied Probability. 26(2):986–1028 (2016).
7. Dettmann CP, Georgiou O.: Random geometric graphs with general connection functions. Physical Review E. 93(3):032313 (2016).
8. Waxman BM. Routing of multipoint connections. IEEE journal on selected areas in communications. 6(9):1617-22 (1988).

Geometric randomization of real networks with prescribed degree sequence

Michele Starnini[1], Elisenda Ortiz[2,3], M.Ángeles Serrano[2,3,4,*]

[1] Data Science Laboratory, ISI Foundation, Via Chisola 5, 10126, Torino, Italy
[2] Departament de Física de la Matèria Condensada, Universitat de Barcelona, Martí Franquès 1, 08028 Barcelona, Spain
[3] Universitat de Barcelona Institute of Complex Systems (UBICS), Universitat de Barcelona, 08028, Barcelona, Spain
[4] ICREA, Pg. Lluís Companys 23, 08010, Barcelona, Spain
* marian.serrano@ub.edu

1 Introduction

The practice of testing hypotheses against a properly specified control case, or null model, is at the heart of the scientific method. In network science [1], null models take typically the form of generative models that produce maximally random graph ensembles given some specific features [2, 3]. Many successful applications include the detection of over-represented motifs in networks [4], the quantification of communities using modularity [5], the detection of rich-club ordering [6, 7] or the characterization of structural correlations in weighted networks [8]. However, null models for networks that incorporate geometric information are scarce and mainly focused on spatial networks.

In fact, a geometric approach to the structure of complex networks has only started to be developed recently. A class of these models in hidden metric spaces [9, 10] explains many pivotal features of real networks simultaneously, including the small world property, heterogeneous degree distributions and high levels of clustering. In those models, the probability of connecting two nodes is determined by their distance in an underlying metric space. This distance is defined along two dimensions representing popularity and similarity features of the nodes, such that the more popular and the more similar two nodes are, the greater the chance to interact and be linked. Specifically, in the well-known $\mathbb{S}^1$ model [9], the hidden degree of a node is a proxy for its popularity, and nodes are assigned angular positions in a circle, such that the angular separation between nodes provides a measure of their similarity. The hidden degree can be reinterpreted as a radial coordinate in a hyperbolic plane [11], leading to the formulation of the isomorphic $\mathbb{H}^2$ model. In both geometric network models, $\mathbb{S}^1$ and $\mathbb{H}^2$, the angular coordinate is uniformly distributed, at odds with the heterogeneous angular distributions observed in hyperbolic maps of real networks [12–14]. In such maps, clusters of nodes lying nearby in the similarity space form indeed geometric communities [13, 14]. This observation opens the door to the use of geometric models with homogeneous angular distribution as null models for the investigation of the community organization and other structural properties of real networks.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Here, we introduce a model for the randomization of complex networks with geometric structure consisting of a rewiring procedure [15] based in the popularity-similarity $\mathbb{S}^1$ network model. The geometric randomization (GR) model, as we named it, preserves exactly the degree sequence of the input network while completely randomizes the angular coordinates of the nodes. Such randomization of the similarity coordinate supports the use of the GR as a null model for the analysis of the topological properties of real networks, including community structure. The GR model assumes the same form of the connection probability as in the $\mathbb{S}^1$ model, and a uniform distribution for the similarity coordinate as well. In contrast, it is fit with a given degree-sequence. Gainfully, the use of prescribed degrees allows to skip the delicate task of estimating hidden degree variables from real data. This attribute can help, for instance, in the analysis of features which are specially sensitive to fluctuations in the degree cutoff, like the behavior of dynamical processes such as epidemic spreading or synchronization, or for high-fidelity reproduction of real network topologies.

2 Results

Based on the premises mentioned above, we propose an algorithm that produces a randomized version of an original network by homogenizing the angular distribution, rewiring the links, preserving the given degrees and maximizing the likelihood that the new topology is generated by the geometric $\mathbb{S}^1$ model. The GR model is manifestly simple as it relies upon a single free parameter, β, controlling the clustering of the rewired replica. Initially, we show how to tune this parameter, and conclude that the parameter value needed for the GR replica to have the same level of clustering as the original network is in general different from the estimated β value of the original network.
Secondly, we analyze the effects of the GR model on the topological properties of 6 real networks from different domains. We show that the deviations between GR and original networks at the level of clustering and average nearest neighbors degree spectrums are almost inexistent. Nonetheless, this is not the case for the replicas of real networks obtained directly from $\mathbb{S}^1$ model through hidden degree estimation. This observation informs the importance of preserving the exact degree sequence during the construction of an angularly randomized version of a network in order to not to alter its main topological characteristics.
Lastly, we demonstrate the applicability of GR by implementing it as a null model for the analysis of community structure. We focus on the comparison of modularity measures obtained using both topological (Louvain method [16]) and geometrical (Critical Gap Method [13]) approaches, while preserving the clustering between original and GR networks. As a result, we find that geometric and topological communities detected in real networks are consistent, while topological communities are also detected in randomized counterparts as an effect of structural constraints. The fact that an underlying geometric organization imposes structural constraints on complex networks, which are strong enough for recreating detectable topological communities even in the absence of geometric ones, is an interesting subject by itself and will be investigated in future work.

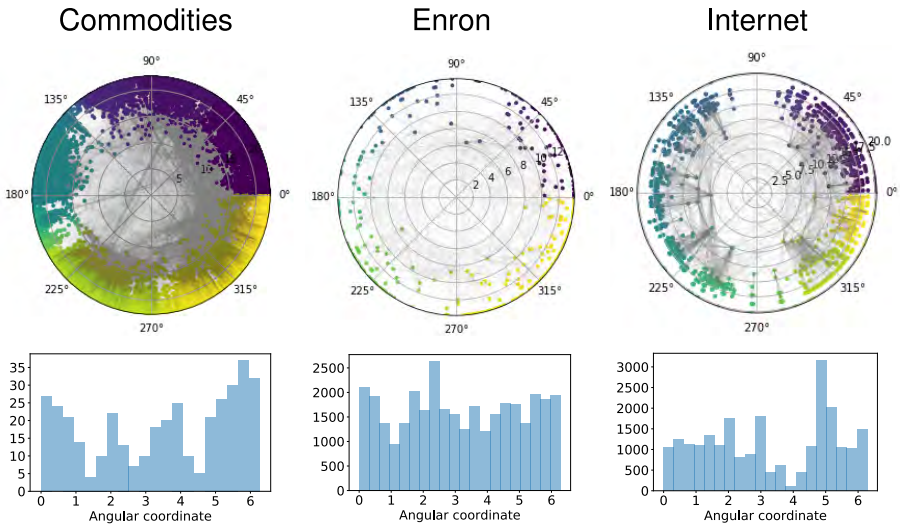


Fig. 1. Top row: Empirical networks embedded in the hyperbolic disk. Distinct communities are indicated by different colors as detected by the Critical Gap Method. **Bottom row**: Probability distribution of the angular coordinate, $P(\theta)$, of the empirical networks.

References

1. Newman M E J 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
2. Bianconi G 2009 *Phys. Rev. E* **79** 036114
3. Cimini G, Squartini T, Saracco F, Garlaschelli D, Gabrielli A and Caldarelli G 2019 *Nat. Rev. Phys.* **1** 5871
4. Shen-Orr S, Milo R, Mangan S and Alon U 2002 *Nat. Genet.* **1** 648
5. Newman M E J and Girvan M 2004 *Phys. Rev. E* **69** 026113
6. Serrano M A 2008 *Phys. Rev. E* **78** 026101
7. Colizza V, Flammini A, Serrano M A and Vespignani A 2006 *Nat. Phys.* **2** 1105
8. Garlaschelli D and Loffredo M I 2009 *Phys. Rev. Lett.* **102** 038701
9. Serrano M A, Krioukov D and Boguñá M 2008 *Phys. Rev. Lett.* **100** 078701
10. Boguñá M, Krioukov D and Claffy K C 2009 *Nat. Phys.* **5** 7480
11. Krioukov D, Papadopoulos F, Kitsak M, Vahdat A and Boguñá M 2010 *Phys. Rev. E* **82** 036106
12. Boguñá M, Papadopoulos F and Krioukov D 2010 *Nat. Comm.* **1** 62
13. Serrano M A, Boguñá M and Sagus F 2012 *Mol. BioSyst.* **8** 84350
14. García-Pérez G, Boguñá M, Allard A and Serrano M A 2016 *Sci. Rep.* **6** 33441
15. Maslov S and Sneppen K 2002 *Science* **296** 9103
16. Blondel V D, Guillaume J L, Lambiotte R and Étienne L 2008 *J. Stat. Mech.* 10008

Small worlds and clustering in spatial networks

Marián Boguñá[1,2], Dmitri Krioukov[3,4] Pedro Almagro[1,2], and M. Ángeles Serrano[1,2,5]

[1] Departament de Física de la Matèria Condensada, Universitat de Barcelona, Martí i Franquès 1, E-08028 Barcelona, Spain
[2] Universitat de Barcelona Institute of Complex Systems (UBICS), Universitat de Barcelona, Barcelona, Spain
[3] Network Science Institute, Northeastern University, 177 Huntington avenue, Boston, MA, 022115
[4] Department of Physics, Department of Mathematics, Department of Electrical & Computer Engineering, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, MA 02115, USA
[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, E-08010 Barcelona, Spain

1 Introduction

In spatial networks, nodes are positioned in a geometric space, and the distances between them in the space affect their linking probability in the network [1]. In real-world systems, such spaces can be explicit/physical, as in geographically embedded networks [2, 3] or in the Ising model with long-range interactions [4–6]. Yet these spaces can be also hidden/latent. Latent similarity spaces have been employed for nearly a century to model homophily in social networks, for instance [7, 8]: the closer the two people are in a virtual similarity space, the more similar they are, the more likely they know each other [9]. Another field where the space can be virtual are graph embeddings in computer science and machine learning, with applications including network compression, visualization, and node labeling [10, 11].

In models of spatial networks, the space is usually explicit. Perhaps the simplest spatial network model is that of random geometric graphs that have been extensively studied in mathematics and physics since the early 60ies [12–15]. In these graphs, nodes are positioned in a space randomly using a point process, usually a Poisson point process, and two nodes are linked in the graph if the distance between them in the space is less than a fixed threshold. If the intensity of the point process does not depend on the graph size n, then the resulting graphs are sparse and have nonzero clustering in the thermodynamic $n \to \infty$ limit, thus sharing these two properties with many real-world complex networks [16, 17]. Yet many of these networks are also heterogeneous small worlds, while random geometric graphs are homogeneous large worlds.

This mismatch was resolved in [18, 19] where a class of models of spatial networks that are sparse heterogeneous small worlds with nonzero clustering was introduced. Networks in these models have some additional properties commonly observed in real-world networks, such as self-similarity [18, 20] and community structure [21–23]. Yet the following question remains: what are the general requirements to spatial network models so that networks in these models possess the properties of real-world networks?

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

Here [24], we first focus on just three properties: (1) sparsity, (2) small worldness, and (3) nonzero clustering. Simplifying the results a bit, we show that spatial networks in $\mathbb{R}^d$ have all these three properties at once only if the probability p_{ij} of connection between nodes i and j scales with the distance x_{ij} between them in $\mathbb{R}^d$ as $p_{ij} \sim x_{ij}^{-\beta}$ with $\beta \in (d, 2d)$. We then add (4) heterogeneity to the list of the requirements, and show that β must be within the same range $(d, 2d)$ if the variance of the degree distribution is finite. If it is infinite, however, e.g. if it is a power law with exponent $\gamma \in (2, 3)$, then the networks are always ultrasmall worlds, and any $\beta > d$ satisfies all the four requirements. Finally, we show that if we also want to suppress nonstructural degree correlations, then the unique shape of the connection probability in the heterogeneous case is as in [18, 19]: $p_{ij} \sim (\kappa_i \kappa_j)^{\beta/d} x_{ij}^{-\beta}$, where κ_i, κ_j are the expected degrees of nodes i, j.

References

1. Marc Barthélemy. Spatial networks. *Phys Rep*, 499(1-3):1–101, 2011.
2. Vito Latora and Massimo Marchiori. Is the Boston subway a small-world network? *Phys A Stat Mech its Appl*, 314(1-4):109–113, 2002.
3. R. Guimera, S Mossa, A Turtschi, and L a N Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc Natl Acad Sci*, 102(22):7794–7799, 2005.
4. M. Gitterman. Small-world phenomena in physics: the Ising model. *J Phys A Math Gen*, 33(47):8373–8381, 2000.
5. J. Viana Lopes, Yu G. Pogorelov, J. M. B. Lopes dos Santos, and R. Toral. Exact solution of Ising model on a small-world network. *Phys Rev E*, 70(2):026112, 2004.
6. Tarcísio N. Teles, Fernanda P. da C. Benetti, Renato Pakter, and Yan Levin. Nonequilibrium Phase Transitions in Systems with Long-Range Interactions. *Phys Rev Lett*, 109(23):230601, 2012.
7. A. P. Sorokin. *Social Mobility*. Harper, New York, 1927.
8. Giandomenico Majone. Social space and social distance: Some remarks on metric methods in data analysis. *Qual Quant*, 6(2), 1972.
9. Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annu Rev Sociol*, 27(1):415–444, 2001.
10. Aditya Grover and Jure Leskovec. node2vec. In *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min - KDD '16*, 2016.
11. Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Syst*, 151:78–94, 2018.
12. E. N. Gilbert. Random Plane Networks. *J Soc Ind Appl Math*, 9(4):533–543, 1961.
13. Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, Oxford, 2003.
14. Jesper Dall and Michael Christensen. Random geometric graphs. *Phys Rev E*, 66(1):016121, 2002.
15. Justin P. Coon, Carl P. Dettmann, and Orestis Georgiou. Entropy of Spatial Network Ensembles. 042319:1–7, 2017.
16. Albert-László Barabási. *Network science*. Cambridge University Press, Cambridge, UK, 2016.
17. M. E. J. Newman. *Networks*. Oxford University Press, Oxford, 2018.

18. M. Ángeles Serrano, Dmitri Krioukov, and Marián Boguñá. Self-Similarity of Complex Networks and Hidden Metric Spaces. *Phys Rev Lett*, 100(7):078701, 2008.
19. Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Phys Rev E*, 82(3):036106, 2010.
20. Guillermo García-Pérez, Marián Boguñá, and M. Ángeles Serrano. Multiscale unfolding of real networks by geometric renormalization. *Nat Phys*, 14(6):583–589, 2018.
21. Konstantin Zuev, Marián Boguñá, Ginestra Bianconi, and Dmitri Krioukov. Emergence of Soft Communities from Geometric Preferential Attachment. *Sci Rep*, 5(1):9421, 2015.
22. Alessandro Muscoloni and Carlo Vittorio Cannistraci. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New J Phys*, 20(5):052002, 2018.
23. Guillermo García-Pérez, M. Ángeles Serrano, and Marián Boguñá. Soft Communities in Similarity Space. *J Stat Phys*, 173(3-4):775–782, 2018.
24. Marián Boguñá, Dmitri Krioukov, Pedro Almagro, and M Ángeles Serrano. Small worlds and clustering in spatial networks. *arXiv:1909.00226*, 2019.

The distribution of shortest path lengths in configuration model networks and other random networks

Ofer Biham[1], Eytan Katzav[1], and Reimer Kühn[2]

[1] Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel,
[2] Mathematics Department, King's College London, Strand, London WC2R 2LS, UK
oferbiham@gmail.com

The theory of complex networks provides a useful conceptual framework for the study of a large variety of systems and processes in science, technology and society. These studies are based on network models, in which the nodes represent physical or virtual objects, while the edges represent the interactions between them. Typically, these networks exhibit random structures, which can be characterized by their statistical properties at the local and global scales. The local structure of a network is captured by the degree distribution and by certain correlations between the properties of nearby nodes. However, the large scale structure of a network is captured by the spectrum of path lengths between random pairs of nodes. The shortest path between each pair of nodes is of particular importance because it provides the strongest interaction and fastest response between these nodes. The average lengths of these paths were studied extensively, while the entire distribution of shortest path lengths (DSPL), $P_{\rm DSPL}(L=\ell)$, has attracted limited attention.

The DSPL provides a useful framework for structural analysis of networks, such as network hyperbolicity [1], as well as for analysis of dynamical processes on networks, such as the propagation of information, traffic navigation [2], and epidemic spreading. To give just one example, considering an epidemic which starts from a random individual, i, in the limit of high infection rate, the temporal spreading of the infection is determined by the shell structure around node i (Fig. 1). Thus, the expectation value of the number of nodes infected up to time t, $N_{\rm I}(t)$, can be expressed in terms of the DSPL and is given by $N_{\rm I}(t) = 1 + (N-1)\,[1 - P_{\rm DSPL}(L > t)]$, where N is the network size.

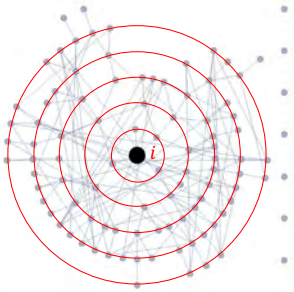


Fig. 1. Illustration of the shell structure around a reference node in a random network. The $\ell^{\rm th}$ shell consists of the nodes which are at distance ℓ from the reference node.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Recently, we developed a suite of analytical approaches for the calculation of the DSPL in a wide range of random networks both in and out of equilibrium. The first approach is the Random Shells Approach (RSA), which is designed to treat Erdős-Rényi (ER) networks [3]. We later improved RSA in Ref. [4] to take into account the detailed size and micro-structure of the giant cluster. This improvement yields very good results even in the vicinity of the percolation transition of the ER ensemble, which takes place when the mean degree is $c = 1$. To obtain a complete understanding of the problem, including the subpercolating regime, we have developed a different methodology, based on a topological expansion [5], which yields an exact result, namely an exponential distribution, for the DSPL with $c < 1$, conditioned on the nodes being on the same cluster. An interesting conclusion is that the mean distance between random nodes is $E[L|L < \infty] = 1/(1-c)$, which means that in the vicinity of the percolation transition it diverges. Among other things, it means that even within the ER ensembles the common lore that distances are "small-world" is far from being the full picture.

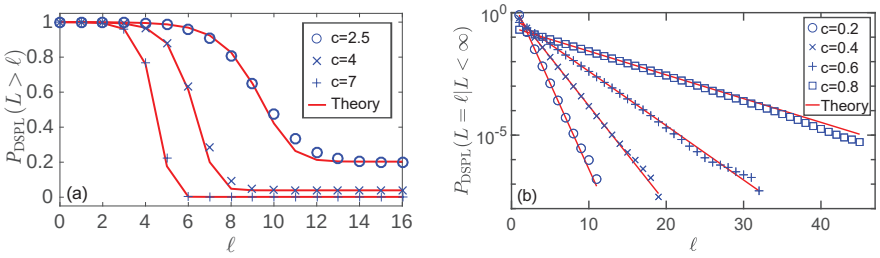


Fig. 2. (a) The tail distribution of shortest path lengths, $P_{\mathrm{DSPL}}(L > \ell)$, for Supercritical ER networks above the percolation threshold, using the Random Paths Approach. (b) The exact result for the DSPL conditioned on finite distances $P_{\mathrm{DSPL}}(L = \ell|L < \infty)$ for Subcritical ER networks below the percolation threshold, using the topological expansion. In both cases $N = 10^4$. The analytical results (solid lines) agree with the results of computer simulations (symbols).

The second methodology, named the Random Paths Approach (RPA), was also developed in the context of the ER ensemble [3], but unlike RSA lent itself to generalization. We identified and formalized the relation between RPA and the cavity method, which allowed application of the RPA to configuration model networks [6], and in particular to random regular graphs (were the exact solution is a Gompertz distribution) as well as to scale-free networks - see Fig. 3. We found that except for the very dilute limit, the distance between most pairs of nodes is centered around to the typical distance (mean or mode), which is given by $\langle L\rangle \simeq \ln N/\ln\left(\langle K^2\rangle/\langle K\rangle - 1\right)$. Also, when the 2$^{\mathrm{nd}}$ moment of the degree distribution diverges (as in certain scale-free networks), the N-dependence enters in a more complicated way, which may lead to an "ultra-small" network [7], i.e. with a mean distance that scales like $\log\log N$ or $\log N/\log\log N$.
Recently, we have developed a methodology to study growing networks based on master equations. We successfully calculated analytically the DSPL for the Node Duplication

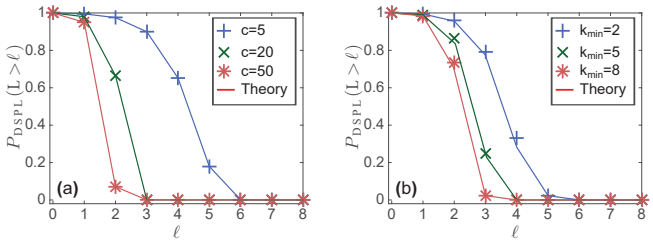


Fig. 3. The tail distributions $P_{\mathrm{DSPL}}(L>\ell)$, for (a) Random Regular Graphs of $N=10^3$ nodes, and degrees $c=5$, 20, and 50, and (b) Scale-Free networks of size $N=10^3$, with $p(k)\propto k^{-2.5}$ and lower cutoffs at $k_{\min}=2$, 5 and 8. The analytical results (solid lines), obtained using the Random Path Approach, compare well to numerical simulations (symbols). Note that larger degrees (a) and $k_{\min}$'s (b) reduce the mean distance of the network.

model, both undirected [8] and directed [9]. It turns out that although networks generated by this model are scale-free their mean distance scales like $\log N$, and therefore they are small-world networks, unlike generic scale free networks, that were shown to be ultra-small [7]. Moreover, the mean distance is even much longer than a corresponding configuration model with the same degree distribution.

The DSPL is only one member in a family of distributions of important metric properties - another one being the distribution of shortest cycle lengths (DSCL). Cycles play an important role in the study of critical phenomena on networks using high temperature expansions as well as in dynamical processes such as the first return of diffusive particles. We calculated the DSCL for the configuration model using the DSPL [10].

References

1. M. Suvakov *et al.*, Hidden geometries of networks arising from cooperative self-assembly, *Sci. Rep.* **8**, 1987 (2018).
2. B. Tadic *et al.*, Information super-diffusion on structured networks, *Physica A* **332**, 566 (2004).
3. E. Katzav *et al.*, Results for the DSPL in random networks, *EPL* **111**, 26006 (2015).
4. I. Tishby, O. Biham, E. Katzav and R. Kühn, Revealing the micro-structure of the giant clusters in random graph ensembles, *Phys. Rev. E* **97**, 042318 (2018).
5. E. Katzav, O. Biham and A. Hartmann, The distribution of shortest path lengths in subcritical Erdős-Rényi networks, *Phys. Rev. E* **98**, 012301 (2018).
6. M. Nitzan, E. Katzav, R. Kühn and O. Biham, Distance distribution in configuration model networks, *Phys. Rev. E* **93**, 062309 (2016).
7. R. Cohen and S. Havlin, Scale-free networks are ultrasmall, *PRL* **90**, 058701 (2003).
8. C. Steinbock, O. Biham and E. Katzav, The DSPL in a class of node duplication network models, *Phys. Rev. E* **96**, 032301 (2017).
9. C. Steinbock, O. Biham and E. Katzav, Exact results for directed random networks that grow by node duplication, arXiv:1807.01591 (2018).
10. H. Bonneau, A. Hassid, O. Biham, R. Kühn and E. Katzav, Distribution of shortest cycle lengths in random networks, *Phys. Rev. E* **96**, 062307 (2017).

Angular separability of data clusters or network communities in geometrical space and its relevance to hyperbolic embedding

Alessandro Muscoloni[1] and Carlo Vittorio Cannistraci[1,2]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy

Abstract

Analysis of 'big data' characterized by high-dimensionality such as word vectors and complex networks requires often their representation in a geometrical space by embedding. Recent developments in machine learning and network geometry have pointed out the hyperbolic space as a useful framework for the representation of this data derived by real complex physical systems. In the hyperbolic space, the radial coordinate of the nodes characterizes their hierarchy, whereas the angular distance between them represents their similarity. Several studies have highlighted the relationship between the angular coordinates of the nodes embedded in the hyperbolic space and the community metadata available. However, such analyses have been often limited to a visual or qualitative assessment. Here, we introduce the angular separation index (ASI), to quantitatively evaluate the separation of node network communities or data clusters over the angular coordinates of a geometrical space [1]. ASI is particularly useful in the hyperbolic space - where it is extensively tested along this study - but can be used in general for any assessment of angular separation regardless of the adopted geometry. ASI is proposed together with an exact test statistic based on a uniformly random null model to assess the statistical significance of the separation. We show that ASI al-lows to discover two significant phenomena in network geometry. The first is that the increase of temperature in 2D hyperbolic network generative models, not only reduces the network clustering but also induces a 'dimensionality jump' of the network to dimensions higher than two. The second is that ASI can be successfully applied to detect the intrinsic dimensionality of network structures that grow in a hidden geometrical space.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

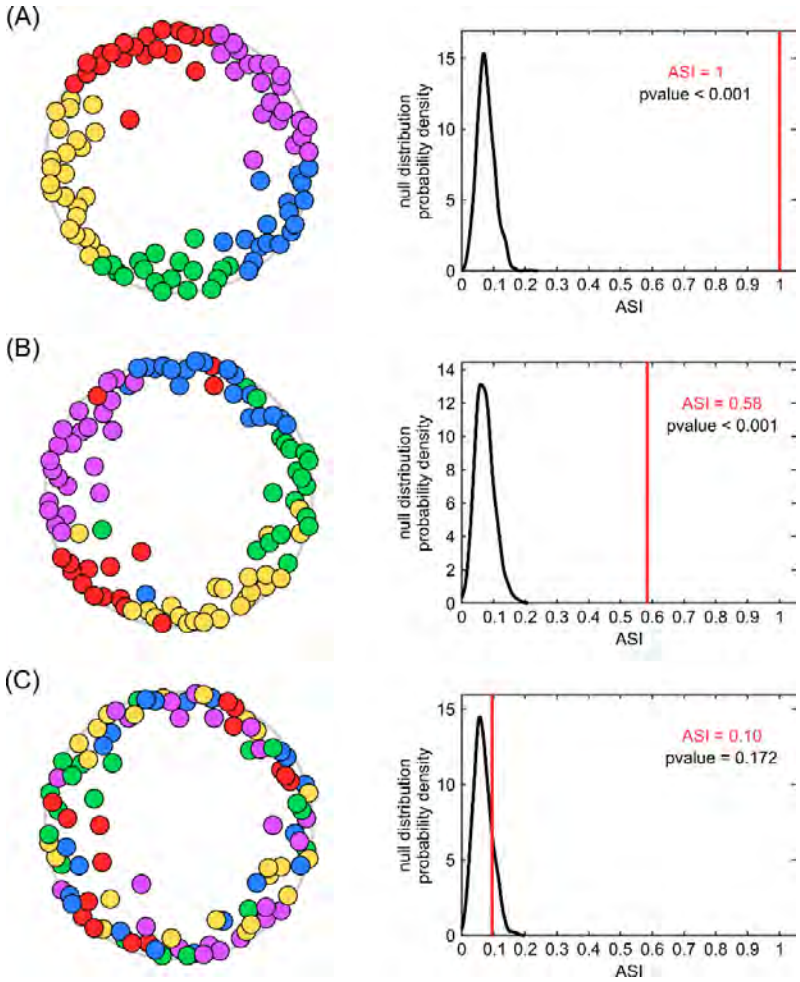


Fig. 1. Angular separation in 2D with statistical test. The left panels show examples of 2D hyperbolic embeddings of synthetic networks generated using the nPSO model. (A) The nPSO network has been generated with parameters $N = 100$ (network size), $m = 3$ (half of average degree), $T = 0.1$ (temperature, inversely related to the clustering coefficient), $C = 5$ (number of communities) and $\gamma = 3$ (power-law degree distribution exponent). The embedded coordinates have been inferred using the coalescent embedding method RA2-ncISO-EA. The 5 ground-truth communities are highlighted with different colors. (B) The nPSO network has been generated with the same parameters as in (A), except for $T = 0.9$. The embedded coordinates have been inferred using the coalescent embedding method RA2-ncISO-EA. (C) The embedded coordinates correspond to the ones in (B) after a random reshuffling. The right panels represent the statistical test for the ASI evaluation and show the observed ASI (in red) compared to the null distribution of ASIs (in black), reporting the related p-value. For further details, please see the reference [1].

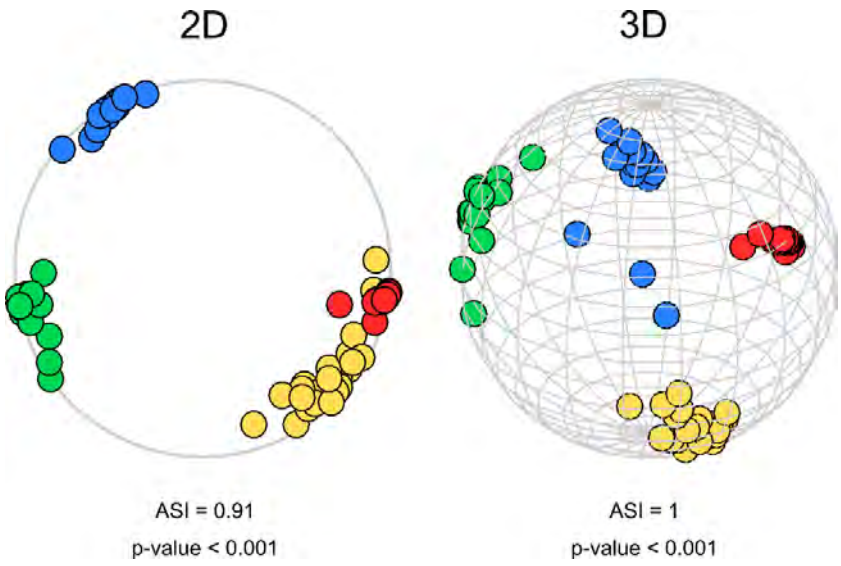


Fig. 2. ASI improvement in 3D with respect to 2D. The figure shows the hyperbolic embedding of the *opsahl_10* network using the coalescent embedding method RA1-ISO both in the 2D hyperbolic disk (left) and in the 3D hyperbolic sphere (right). The 4 ground-truth communities are highlighted with different colors. At the bottom of each panel the ASI and the related p-value of the statistical test are reported. The figure provides an example in which the addition of the third dimension of embedding improves the angular separation of the communities, leading to a perfect segregation. For further details, please see the reference [1].

References

1. A. Muscoloni, and C. V. Cannistraci, “Angular separability of data clusters or network com-munities in geometrical space and its relevance to hyperbolic embedding”, arXiv:1907.00025, 2019.

Part XII

Network Models

Structure of the giant component and statistics of articulation points in configuration model networks

Ido Tishby[1], Eytan Katzav[1], Ofer Biham[1], and Reimer Kühn[2]

[1] Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel,
[2] Mathematics Department, Kings College London, Strand, London WC2R 2LS, UK
ido.tishby@mail.huji.ac.il

Network models provide a useful conceptual framework for the study of a large variety of systems and processes in science, technology and society. One of the central lines of inquiry in the study of random networks, has been concerned with the existence, under suitable conditions, of a giant component. Critical parameters for the emergence of a giant component in Erds-Rényi (ER) networks were identified and its asymptotic size was determined ([1]). For configuration model networks, those problems were solved by Molloy and Reed ([2]). However, not much work has been done concerning the finer structure of the giant component, which is the main aim of [3]. The knowledge of degree distributions and degree-degree correlations restricted to the giant component of a network would be very useful when investigating dynamical processes on complex networks. For example, epidemic spreading is usually studied in restriction to the giant component on which the contamination is potentially maximal ([4]).

For networks in the configuration model class, we obtain the degree distribution $P(k|1)$, conditioned on the giant component, as a relation to the degree distribution $P(k)$ of the entire network. In Fig. 1 we compare these two distributions for an ER network (left) and for a scale-free configuration model network (right). As can be seen, the analytical results agree with the simulations.

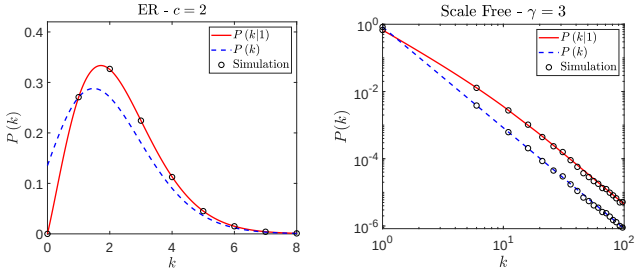


Fig. 1. Analytical results for the degree distribution, $P(k|1)$, of the giant component of an ER network with $c = 2$ (left) and a scale-free network with $\gamma = 3$, $k_{max} = 100$ (right), as well as results of computer simulations (circles). For comparison, the degree distributions, $P(k)$, of the whole networks are also shown (dashed lines).

A consequence of the relations obtained between $P(k)$ and $P(k|1)$, is that they can be inverted, so to prescribe the degree distribution of the entire network given that of the giant component. These inverted relations were obtained in [5], and are employed as a

method for the construction of ensembles of random networks that consist of a single connected component with a desired degree distribution. This approach extends the construction toolbox of random networks beyond the configuration model framework, in which one controls the degree distribution but not the number of components and their sizes.

In a second broader consequence of [3], we address in [6] a closely related topic: the structural vulnerability of the giant component and the network as a whole. Networks are often exposed to the loss of nodes and edges, which may severely affect their functionality. Such losses may occur due to inadvertent node failures ([7]), propagation of epidemics or deliberate attacks ([8]). In each network, one can identify the nodes whose deletion would break the component on which they reside into two or more components. Such nodes are called articulation points (APs) or cut vertices. Sometimes the resulting disruption caused by their removal can be so severe as to cause the entire system to fail, making such APs essentially single point failures (SPOF) of the system. In this context, a concept of a strongly connected component (SCC) emerges, which is the maximal sub-component of the network that does not contain any APs with respect to itself. In Fig. 2 we present a schematic illustration of an ER network in which these different concepts are visualized.

We derived analytical results for the statistical properties of APs in ER networks and configuration model networks with various degree distributions. We obtain the probability $P(i \in \mathrm{AP})$ that a random node i is an AP, and calculate various related conditional probabilities, such as $P(i \in \mathrm{AP}|k)$ and $P(i \in \mathrm{AP}|\mathrm{GC})$, relating the probability that a node is an AP, to its degree k, or to the probability that it resides on the giant component of the network, respectively. We also introduce a new AP-based centrality measure: We denote by r the number of components which are added to the network upon deletion of a given node i, and refer to this as the articulation rank of this node. We obtain analytical results for the distribution of articulation ranks, $P(R = r)$ of all the nodes in the network. In Fig. 3 we show (left) $P(R = r)$ for an ER network of mean degree $c = 2$, and also the mean articulation rank $\langle R \rangle$ (right) as a function of mean degree c in ER networks. The theoretical curves agree well with the simulations.

References

1. B. Bollobás, Random graphs (Cambridge University Press, 2001)
2. M. Molloy and A. Reed, The size of the giant component of a random graph with a given degree sequence, *Combin., Prob. and Comp.* **7** , 295-305 (1998)
3. I. Tishby, O. Biham, E. Katzav and R. Kühn, Revealing the Micro-Structure of the Giant Component in Random Graph Ensembles, *Phys. Rev. E.* **97**, 042318 (2018)
4. M. Newman, Networks, 2nd Ed. (Oxford University Press, 2018)
5. I. Tishby, O. Biham and E. Katzav, Generating random networks that consist of a single connected component with a given degree distribution, *Phys. Rev. E.* **99**, 042308 (2019)
6. I. Tishby, O. Biham, R. Kühn and E. Katzav, Statistical analysis of articulation points in configuration model networks, *Phys. Rev. E.* **98**, 062301 (2018)
7. R. Cohen, K. Erez, D. ben-Avraham and S. Havlin, Resilience of the internet to random breakdowns, *Phys. Rev. Lett.* **85**, 4626-4628 (2000)
8. R. Cohen, K. Erez, D. ben-Avraham and S. Havlin, Breakdown of the internet under intentional attack, *Phys. Rev. Lett.* **86**, 3682-3685 (2001)

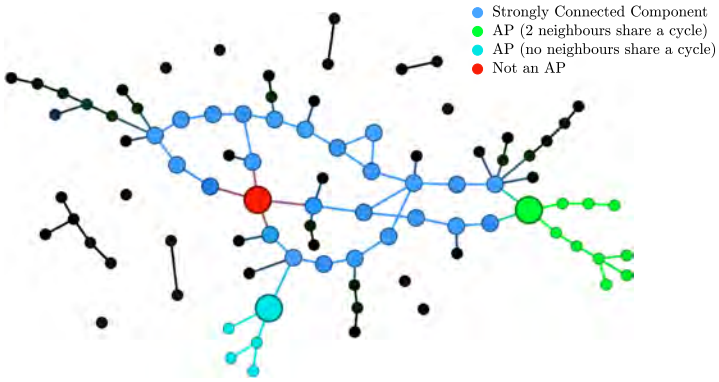


Fig. 2. Illustration of an ER network, with some structural features emphasized. In blue are marked the nodes residing on the strongly connected component of the network. 3 nodes, with distinct surrounding environments are shown in large circles. In cyan is marked an articulation point (AP) of degree $k = 3$ in a tree component. Deletion of the AP would split the giant component into three separate components; In green is marked an AP of degree $k = 4$, where two of its neighbours reside on a cycle. Deletion of the AP would split the giant component into three separate components; The portions that break off from the giant component upon the removal of each AP are coloured the same as the APs. Lastly the node marked in red is not an AP because each pair of its neighbours share a cycle. As a result, upon deletion of the red node all its neighbours remain on the giant component.

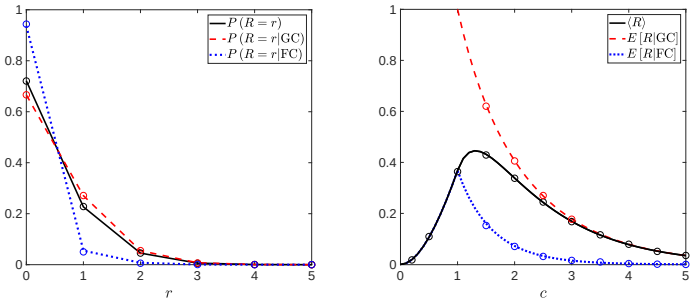


Fig. 3. Analytical results for the distribution $P(R = r)$ (left) of the articulation ranks of nodes in an ER network with $c = 2$ (solid line) and the mean articulation rank $\langle R\rangle$ (right) of all nodes in the network. These quantities are also shown for nodes restricted to the giant component (dashed line), and nodes in the finite components (dotted line). The analytical results are in very good agreement with the results of computer simulations (circles).

Complex distributions emerging in compression and filtering

Gareth J. Baxter[1], Rui A. da Costa[1], Sergey N. Dorogovtsev[1,2], and José F. F. Mendes[1,3]

[1] Departamento de Física da Universidade de Aveiro & I3N,
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal,
sdorogov@ua.pt,
WWW home page: http://sweet.ua.pt/sdorogov/
[2] A.F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia
[3] School of Computer and Communication Sciences and School of Life Sciences,
École Polytechnique Fédéral de Lausanne, 1015 Lausanne EPFL, Switzerland

1 Introduction

In filtering, each output is produced by a certain number of different inputs. This number is the output's degeneracy. A similar problem emerges in cooperative systems with a large number of local minima in the energy landscape, in particular, in spin glasses and deep learning neural networks. The configuration space of a system of this sort can be divided into a set of domains (basins) of attraction of these minima. One can ask: what is the statistics of these domains of attraction, what is the distribution of their sizes? These issues were explored in a recent series of works [1–4] which exploited the principle of maximum entropy in application to compression problems. The finding of Refs. [1–4] is that optimal compression generates outputs with broad distributions. More specifically, the entropy optimization based theory predicts power-law like distributions of degeneracy of maximally informative outputs (minimal sufficient representations).

We explore the statistics of this degeneracy in an explicitly treatable filtering problem in which filtering performs the maximal compression of relevant information contained in inputs [5]. The filter patterns in this problem conveniently allow a microscopic, combinatorial consideration. This allows us to find the statistics of outputs, namely the exact distribution of output degeneracies, for relatively large input sizes and to describe the dependence of this distribution on the input size and on the size of the input data set.

2 Filtering problem

Let the input data be a set of N strings of zeros and ones of length n, assuming the periodic condition. We consider two types of data set. The first set is the complete set of all possible unique inputs. Its size N is determined by the size n of inputs, $N = 2^n$. Second, we consider data sets of arbitrary size N consisting of strings of uniformly randomly generated zeroes and ones constrained by the same periodic condition.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The filter works as follows: every instance of a specific pattern in the input is marked by a one in the corresponding position in the output. All other positions are marked with zeros. This produces a minimal coding of the positions of the pattern occurrences in the input. For the sake of simplicity, we use the following filter. Each sequence of ones of length 1 in the input (i.e., every one whose neighbors are both zeros) gives one at the same position in the output. All other sequences of ones or zeros in the input produce zeros in the corresponding places in the output.

3 Results

Figure 1(a) represents the statistics of the outputs generated by the complete input data set, 2^n inputs, $\mathcal{N}(d)$ is the number of outputs of degeneracy d, $\mathcal{N}_{\rm cum}(d) = \sum_{u \geq q} \mathcal{N}(q)$.

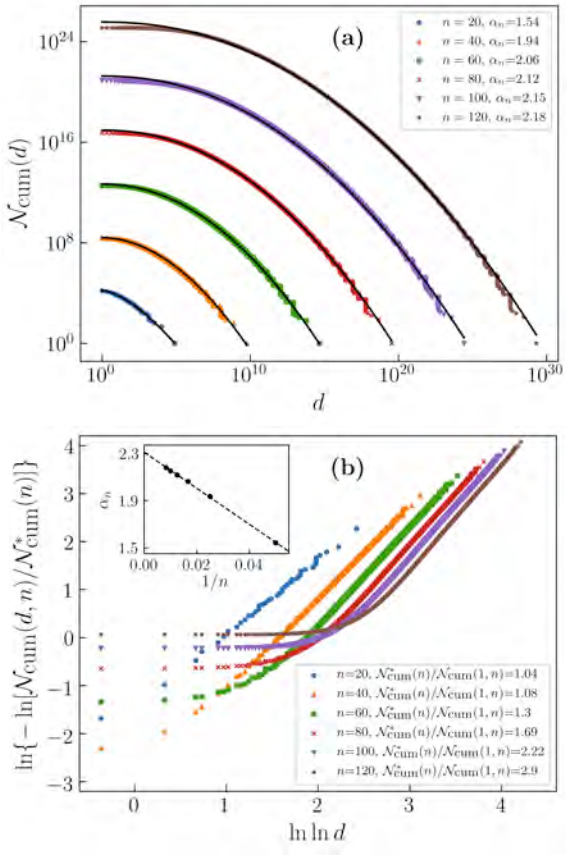


Fig. 1. (a) Cumulative degeneracy distribution for $n = 20, 40, 60, 80, 100, 120$. The black curves represent least-squares fittings of $\ln \mathcal{N}_{\rm cum}(d,n)$ as $\ln \mathcal{N}^*_{\rm cum}(n) + B_n \ln^{\alpha_n} d$ for each n. (b) Cumulative degeneracy distribution $\ln\{-\ln[\mathcal{N}_{\rm cum}(d,n)/\mathcal{N}^*_{\rm cum}(n)]\}$ vs. $\ln\ln d$ for $n = 20, 40, 60, 80, 100, 120$. Inset: exponent α vs. $1/n$.

Figure 1(b) demonstrates that, asymptotically,

$$\mathscr{N}_{\text{cum}}(d) \propto e^{-c\ln^{\alpha} d} = d^{-c\ln^{\alpha-1} d}, \tag{1}$$

where exponent α depends on n, approaching approximately 2.3 as $n \to \infty$. This is significantly distinct from a power-law dependence. Note that the explored range of degeneracies 30 orders of magnitude which enables us to extract the complicated asymptotic dependence, Eq. (1). Each point in Figure 1 was obtained exactly. Importantly, the statistics of outputs is determined not by the form of filter patterns but rather by what occurs in the gaps between them. The degeneracy corresponding to each such gap can be found using recursion relationships. We then used an integer partitions apparatus to aggregate the statistics of prime degeneracies from these gaps, finding the exact full spectrum of output degeneracies.

We inspected the dependence of the form of the degeneracy distributions on the input size n and on the size N of the uniformly randomly generated input data set. These size effects turn out to be very different from those for more familiar distributions drawn from heavy tailed ones, e.g., power-law degree distributions [6]. Curiously, the distributions found for different values of n have a very similar form for input sizes N chosen such that $(z_d/2)^n N$ is constant.

Summary. Our straightforward, purely combinatorial treatment reveals features of distributions of outputs hidden from other approaches. For complete input data sets passed through our filter, we have obtained degeneracy distributions markedly distinct from power laws. Our model filter can be used as a convenient reference filtering and compression problem.

References

1. Cubero, R., Marsili, M, Roudi, Y.: Minimum description length codes are critical. Entropy 20, 755 (2018)
2. Haimovici, A., and Marsili, M.: Criticality of mostly informative samples: A Bayesian model selection approach. J. Stat. Mech.: Theory and Experiment 2015, P10013 (2015)
3. Cubero, R. J., Jo, J., Marsili, M., Roudi, Y., and Song, J.: Minimally sufficient representations, maximally informative samples and Zipfs law. arXiv:1808.00249 (2018)
4. Song, J., Marsili, M:, Jo, J.: Resolution and relevance trade-offs in deep learning. J. Stat. Mech.: Theory and Experiment 2018, 123406 (2018)
5. Baxter, G. J., da Costa, R. A., Dorogovtsev, S. N., Mendes, J. F. F.: Complex distributions emerging in filtering and compression. arXiv:1906.11266 (2019)
6. Dorogovtsev, S. N., J. Mendes, J. F. F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press (2003) van Mieghem, P.: Graph Spectra for Complex Networks. Cambridge University Press (2010)

Field theory for recurrent mobility

Mattia Mazzoli[1], Alex Molas[1], Aleix Bassolas[1], Maxime Lenormand[2], Pere Colet[1], and Jose J Ramasco[1]

[1] IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain
mattia@ifisc.uib-csic.es,
home page: https://ifisc.uib-csic.es/en/people/mattia-mazzoli/
[2] Irstea, UMR TETIS, 500 rue JF Breton, 34093 Montpellier, France

1 Introduction

Human mobility has been studied for decades due to the relevant role it plays in a wide spectrum of applications including economic questions and living conditions, city structure, epidemics forecasts, infrastructures design, urban pollution and air quality. Aggregating the home-work trips of individuals, one can compute the so-called Origin-Destination (OD) matrices which for every pair (i, j) collect the flow of people traveling from i to j, T_{ij}. These matrices are essential for transport planning since they mathematically encode trip demand. Census and dedicated surveys have dominated the area in terms of mobility data collection until a few years ago. With the rise of big data sources, the availability of large-scale quick-updated data has notably increased. Passive sources such as mobile phone records or GPS-located messages in online social networks have been employed to study mobility and, in particular, to extract OD matrices (see also the recent reviews [1, 2]). The quality of the OD matrices obtained from these new information and communication technologies (ICT) sources has been proven consistent with respect to those provided by surveys in urban areas at spatial scales larger than one square kilometer [3]. The availability of this new data opens the door to tackle and revisit relevant theoretical aspects concerning mobility flows that could not be boarded before. Several models have been proposed to obtain the flows from basic variables as the population. The bet is high since determining transport demand is fundamental for infrastructure building and urban planning. Among all models, two competing frameworks have been used for almost 80 years to characterize mobility flows: the gravity [4, 5] and the intervening opportunity [6, 7] models. Briefly, in the gravity model the flows decay with a certain deterrence function, e.g. with an exponential or power law-like forms, while the intervening opportunity models rely on the ”opportunities”, intended as jobs, found within a given area. A few years ago, the radiation model has been introduced as a physical adaptation of the intervening opportunity concept where the density of opportunities is related to the population [8, 9]. In 1947 a visionary study explored the possibility of defining a scalar potential to describe human mobility [10], but the lack of reliable data hindered further research in this direction.

2 Results

In this work, we introduce a new approach based on the observation that daily commuting flows can be represented as vectors pointing from the origin to the destination, and

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

that these elementary vectors can be summed to produce a mesoscopic vector field. A particular pattern is observed in all the cities under study since the vector field clearly points to the city geographical center. This pattern is illustrated for London in Figure 1a with ODs from Twitter data. This vector field fulfills the Gauss (divergence) theorem and also its rotational is nearly zero in all the space. Note that we are studying empirical information, hence these results are far from trivial and they reveal intrinsic features of aggregated daily human mobility. The existence of a well-behaved mesoscopic field is confirmed with both data from Twitter and census for large urban areas. The first feature allows us to admit that the field is generated by a source and it allows us to study the flux around different closed perimeters. To do this we used essentially circles of different radiuses around the center of the cities. The classical models to reproduce OD matrices, i.e. gravity and radiation model, are then employed to generate fields and their results are tested against the empirical ones. The flux produced by a gravity model with an exponentially decaying deterrence function with the distance fits much better than the radiation model. Additionally, the observed irrotationality of the field allows us to define a scalar potential in the space for each city which shapes the urban commuting mobility of inhabitants. The maximum of the potential is typically located in the center of the cities and it decays as one gets further. This potential is a tool that will crucially contribute to controversial issues such as the functional definition of city limits [11], e.g. in the areas of influence of different cities as it can be seen in the case of the Manchester-Liverpool conurbation (Figure 1b), and the presence of polycenters [12]. The results of this work are available on Nature Communications [13].

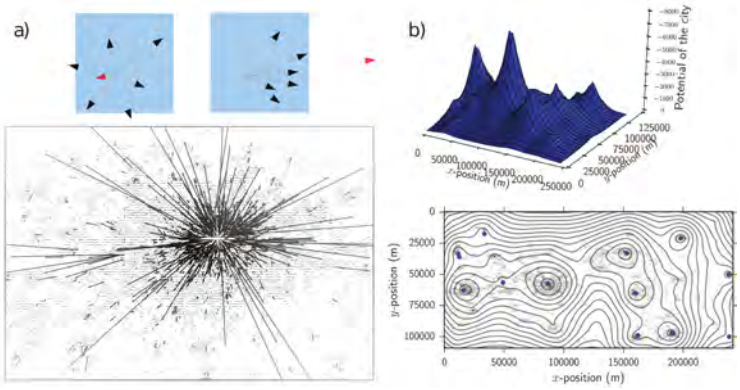


Fig. 1. a) Top row, two examples with the definition of the average vector in every cell (red vector). In the bottom, the vector field in an area comprehending the Greater London area. b) The potential field calculated using the gravity model with an exponential deterrence function in the area of Manchester and Liverpool. We find 13 centers (local maxima).

Summary. We have introduced a vectorial field framework to characterize human mobility flows. When considering recurrent home-work mobility in cities, we find that the mesoscopic field representing the flows is well-behaved in the sense of satisfying

Gauss's theorem and, besides, it is irrotational. As a consequence of this last point, it is possible to define a scalar potential, which reducing the dimensionality of the system encodes all the information on the commuting at a mesoscopic scale. The results are corroborated using two independent data sources for the commuting. The shape of the potential sheds new light on the spatial organization of mobility in cities as we can picture city centers as the strongest gravitational attractors of the metropolitan area and redefine city boundaries. This can have an important practical relevance when planning infrastructures and public services.

References

1. Blondel, D. V., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ Data Science 4, 10 (2015).
2. Barbosa-Filho, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., Tomasini, M.: Human mobility: Models and applications. Physics Reports 734, 1–74 (2018).
3. Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frias-Martinez, E., Ramasco, J. J.: Cross-checking different sources of mobility information. PLoS ONE 9, e105184 (2014).
4. Carey, H. C.: Principles of Social Science, volume 3. JB Lippincott & Company, Philadelphia PA, USA (1867).
5. Zipf, G. K.: The p1 p2/d hypothesis: on the intercity movement of persons. American Sociological Review 11, 677–686 (1946).
6. Stouffer, S. A.: Intervening opportunities: a theory relating mobility and distance. American Sociological Review 5, 845–867 (1940).
7. Ruiter, E. R.: Toward a better understanding of the intervening opportunities model. Transp. Res. 1, 47–56 (1967).
8. Simini, F., González, M. C., Maritan, A., Barabási, A.-L.: A universal model for mobility and migration patterns. Nature 484, 96–100 (2012).
9. Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C., Toroczkai, Z.: Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat. Commun. 5, 5347 (2014).
10. Steward, J. Q.: Empirical mathematical rules concerning the distribution and equilibrium of population. American Geographical Society 37, 461–485 (1947).
11. Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., Batty, M.: Constructing cities, deconstructing scaling laws. Journal of The Royal Society Interface 12, 20140745 (2015).
12. Barthelemy, M.: The Structure and Dynamics of Cities: Urban Data Analysis and Theoretical Modeling. Cambridge University Press, Cambridge, UK (2017).
13. Mazzoli, M., Molas, A., Bassolas, A., Lenormand, M., Colet, P., Ramasco, J. J.: Field theory for recurrent mobility. Nat. Commun. 10, 3895 (2019).

Simulation of virtual networks as excitable media: a particular case of a small-world structure

Yuri Yu. Tarasevich, Alexei V. Rybakov, and Anastasiya V. Mikhailova

Astrakhan State University, Astrakhan, Russia,
tarasevich@asu.edu.ru,
WWW home page: http://science.asu.edu.ru/index.php/user/1/

1 Introduction

Dissemination of ideas and opinions in virtual networks, e.g., in social networks, in the Internet, in academic networks, in the blogosphere, etc. can be considered as excitations in active media [1, 4, 3]. One of the models of active media is the model by Wiener and Rosenblueth [6]. Originally, the model has been proposed to describe conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle. Excitable elements have been considered as vertices of a regular graph (a square lattice). After a while, the model has been generalized [6, 7, 2]. Recently, the generalized Wiener—Rosenblueth model of excitable medium has been used to simulate activities in networks with topologies both of a complete graph [4] and of a scale-free network [3]. The first one can be treated as a small group where everyone knows everyone. The second one can be considered as a model of a virtual group. However, other topologies, e.g. a small-world [5], are possible. Since different networks have different structures, application of the generalized Wiener—Rosenblueth model to other topologies looks promising.

Within the generalized Wiener—Rosenblueth model, each element of the network have three possible states, viz., rest, excitation, refractoriness [6, 7, 2]. Initially, all elements are in the rest state. The i-th element becomes excited under the influence of an external excitation. The intensity of this external excitation must exceed the threshold value, h_i. The i-th element states in the excited state during the time τ_i^e, then it transfers into the refractory state in which it states during the time τ_i^r, then it comes back to the state of the rest. A state of any element is specified by the integer phase, Φ_i^n, and the activator concentration, u_i^n, where integer n indicates the discrete time step. Activator decays with time, g_i is the rate of decay.

Any transitions between these states obey the following set of rules

$$\Phi_i^{n+1} = \begin{cases} \Phi_i^n + 1, & \text{when } 0 < \Phi_i^n < \tau_i^e + \tau_i^r, \\ 0, & \text{when } \Phi_i^n = \tau_i^e + \tau_i^r, \\ 0, & \text{when } \Phi_i^n = 0 \text{ and } u_i^{n+1} < h_i, \\ 1, & \text{when } \Phi_i^n = 0 \text{ and } u_i^{n+1} \geqslant h_i. \end{cases} \tag{1}$$

A vertex receives a certain amount of activator from its adjacent active vertices of the network.

$$u_i^{n+1} = g_i u_i^n + \sum_j I_j^n, \tag{2}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where j is the number of the neighbour vertices i,

$$I_j^n = \begin{cases} 1, & \text{when } 0 < \Phi_j^n \leqslant \tau_j^e, \\ 0, & \text{when } \tau_j^e < \Phi_j^n \leqslant \tau_j^e + \tau_j^r \text{ or } \Phi_j^n = 0. \end{cases} \tag{3}$$

Networks may be both homogeneous and inhomogeneous. Each vertex of a homogeneous network is described by the same set of parameters, while each vertex of a inhomogeneous network has a particular set of parameters. When the vertices are considered as persons, different sets of parameters correspond to persons of different temperaments.

2 Results

A homogeneous small-world network is considered as a model of a virtual community. Activity of the vertices is described by Eqs. (1), (2), and (3). Initially, all vertices are in the rest. Then, some vertices are transferred in an excited state by a sufficient amount of the activator. These excited vertices may activate the whole network. Different regimes can be observed.

Figure 1 demonstrates an example of a periodical regime in a small-world network with 100 vertices. Initially only one vertex got enough amount of activator to transfer into excited state. The scatter plot "min degree" corresponds to the vertex with minimal number of connections, while "max degree" corresponds to the vertex with maximal number of connections. Presented regime is insensitive to choice of initially excited vertex. The parameters of the model are $h = 0.75$, $g = 0.75$, $\tau_e = 5$, $\tau_r = 10$.

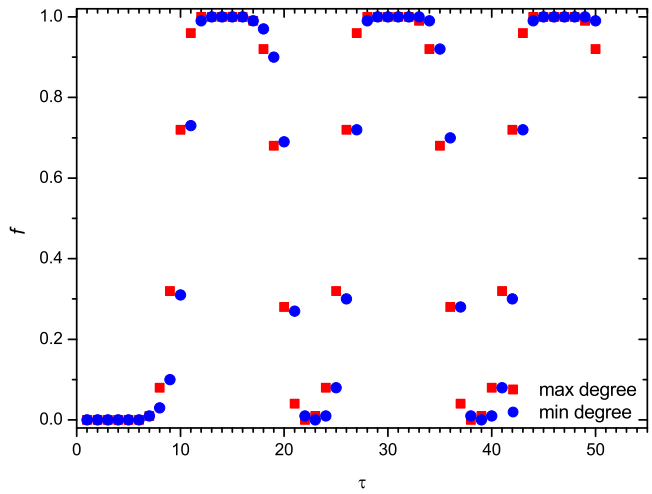


Fig. 1. Example of fraction of excited nodes, f, vs time step, τ. Initially only one node got enough amount of activator to transfer into excited state.

Figure 2 demonstrate propagation of the excitation in the same network.

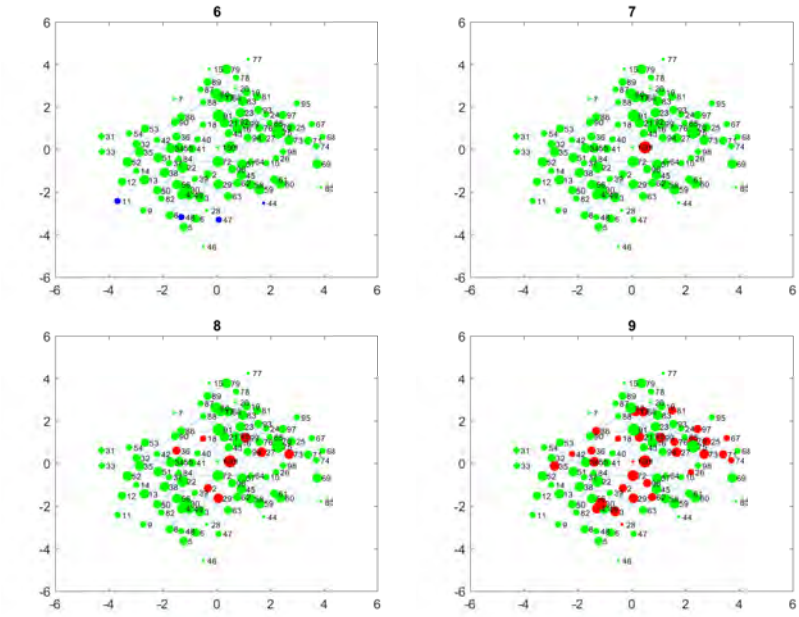


Fig. 2. Propagation of the excitation in the same network $\tau = 6, 7, 8, 9$.

Summary. Dissemination of an activity in a virtual group has been simulated using the generalized model of excitable medium by Wiener—Rosenblueth [6, 7, 2]. A structure of the virtual group has been assumed to be a small-world [5].

References

1. Klimek, P., Bayer, W., Thurner, S.: The blogosphere as an excitable social medium: Richter's and Omori's law in media coverage. Physica A **390**(21–22), 3870–3875 (2011). https://doi.org/10.1016/j.physa.2011.05.033
2. Mikhailov, A.S.: Foundations of Synergetics I: Distributed active systems. Springer series in synergetics, Springer (1990). https://doi.org/10.1007/978-3-642-78556-6
3. Shinyaeva, T.S., Tarasevich, Y.Y.: Virtual network as excitable medium. Journal of Physics: Conference Series **681**, 012008 (feb 2016). https://doi.org/10.1088/1742-6596/681/1/012008
4. Tarasevich, Y.Y., Zelepukhina, V.A.: Academic network as excitable medium. Comp. Res. Model. **7**(1), 177–183 (2015). https://doi.org/10.20537/2076-7633-2015-7-1-177-183, in Russian
5. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (jun 1998). https://doi.org/10.1038/30918
6. Wiener, N., Rosenblueth, A.: The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle. Arch. Inst. Cardiol. Mex. **16**(3), 205–265 (Jul 1946)
7. Zykov, V.S., Mikhailov, A.S.: Rotating spiral waves in a simple model of excitable medium. Sov. Phys. — Doklady **31**, 51–55 (1986)

Analysis of scale-free networks with generalized thresholding functions within the framework of hidden variable formalism

Sámuel G. Balogh[1], Péter Pollner[2], and Gergely Palla[2]

[1] Dept. of Biological Physics, Eötvös University, H-1117 Budapest, Hungary
balogh@hal.elte.hu,
[2] MTA-ELTE Statistical and Biological Physics Research Group, Hungarian Academy of Sciences, H-1117 Budapest, Hungary

1 Introduction

The network approach has become an ubiquitous tool for analyzing complex systems, possibly composed of many interacting sub-units, ranging from the microscopic level to the level of society. In the past few decades one of the most widely studied features of complex networks has been given by the scale-free (SF) property, manifesting in a power-law like decay of $p(k) \sim k^{-\gamma}$. Several growing mechanism have been proposed to generate networks with SF property, such as the well-known Barabási-Albert model together with its modifications and extensions [1, 2]. However, it turned out that, in some cases neither preferential attachment nor the growing mechanisms are available and the topology of the networks are entirely encoded in some hidden hidden properties of the nodes. Inspired by this recognition the concept of static networks with intrinsic node weights (fitnesses, hidden variables) was first investigated by Caldarelli et al. [3] in order to give explanation for the emergence of non-growing scale-free networks. Later on, Boguña et al. [4] proposed a rigorous analytical framework for classes of hidden variable network models and since then the applicability of the model has been confirmed in a large scale ranging from temporal networks, through multifractal networks [5] to hyperbolic networks [6].

2 The non-geographical threshold model

The concept of hidden variables is based on assigning a hidden parameter x_i to each element of a given set of nodes according to an arbitrary but prescribed $\rho(x_i)$ probability density function and then connecting these nodes with $0 \leq f(x_i, x_j) \leq 1$ probability.

It can be shown that SF networks can easily be generated through power-law distribution of fitness, however a surprising result was investigated in [3] where an exponential distribution of fitness ($\rho(x) \sim e^{-x}$) was chosen together with a $f(x,y) = \Theta(x+y-\Delta)$ threshold linking form, where $\Theta(x)$ denotes the Heaviside step function and Δ is a constant (non-geographical threshold model). Under these settings a power law decay of the degree distribution was detected with a $\gamma = 2$ scaling exponent, providing the first evidence that SF networks can be generated in this approach even with non power-law like fitness distributions.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

3 Results

In our work [7] we have analitically extended the results of the non-geographical model ($\gamma = 2$ scaling exponent), now being valid for three different classes of fitness distributions. These generalized classes of fitness distributions and the corresponding linking functions can be written in the following forms:
1, Exponential-like class:

$$\rho_1(x) = H'(x)\exp[-H(x)] \text{ with } f_1 = f_\Delta(H(x)+H(y)) \tag{1}$$

2, Power-like class:

$$\rho_2(x) = G'(x)G^{-\alpha}(x) \text{ with } f_2 = f_\Delta(G(x)G(y)) \tag{2}$$

3, Mixed class:

$$\rho_3(x) = \frac{M'(x)}{(1+M(x))^\alpha} \text{ with } f_3 = f_\Delta(1+M(x)+M(y)+M(x)M(y)) \tag{3}$$

where H, G, M are arbitrary functions while f_Δ is also an arbitrary function but with a lower-cutoff at Δ. The previously defined pairs of functions universally lead to the emergence of SF networks with $\gamma = 2$ thus providing a far-more general form of the non-geographical threshold model. Despite the universality of the exponent networks generated via different forms of f_Δ might show different behaviour at the level of local network quantities such as degree correlation or clustering coefficient. For illustration in Fig. 1. we provide simulation results for the clustering coefficient as a function of k, when replacing the Heaviside step function with other possible forms of f_Δ.

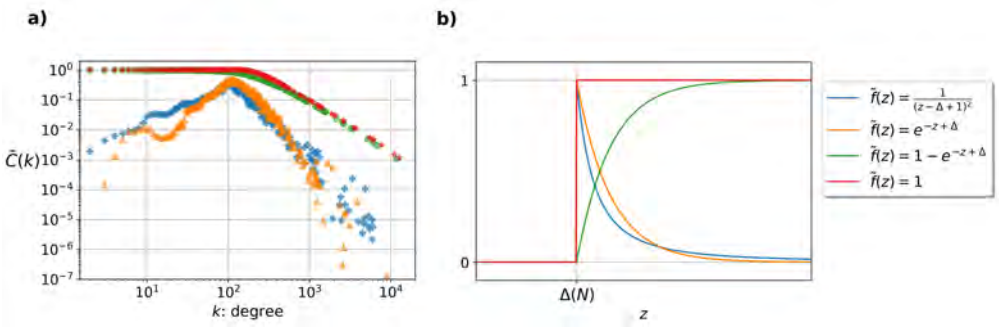


Fig. 1. a) Clustering coefficient as a function of node degrees for four different networks each of them containing $N = 20000$ nodes. All networks were generated by using the same exponential fitness distribution but with different connection functions f_Δ indicated in panel b).

By introducing an external, tuneable β parameter of the linking function (which can be associated with the modulation of the threshold function) $f = f_\beta(\dots)$ similarly to [6],

the scaling exponent of the degree distribution can be modified to $\gamma = \gamma(\beta)$ for each and every particular pair of the above discussed generalized $\{(\rho, f)\}$ classes. Furthermore, we have fully described the $\gamma(\beta)$ transition (Fig. 2.) and also generally discussed the criteria in multiple different cases of how to generate networks having degree distribution independent of the size [7]. Hence, these models with the relaxed threshold at finite β values offer a quite flexible fitness-based approach being applicable to fitness/activity driven systems (such as temporal network or hyperbolic networks) where the distribution of hidden parameters follow non-trivial/complicated distributions.

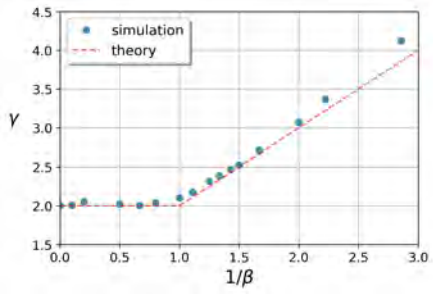


Fig. 2. Scaling exponent γ of the degree distribution as a function of the effective temperature $1/\beta$ in the model with soft thresholding.

From the theoretical point of view it might also be remarkable that according to these results a general mapping can be established between ρ and f, meaning that for any fitness distribution ρ^* there always exists classes of linking functions f^* in the forms of (1),(2),(3) with a lower-cutoff which generate scale-free networks with $\gamma = 2$ scaling exponent and vice versa.

References

1. A.-L. Barabási and R. Albert: Emergence of scaling in random networks. Science, 286:509-512, (1999).
2. G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. Europhysics Letters, 54: 436-442, (2001).
3. G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz. Scale-free networks from varying vertex intrinsic fitness. Phys. Rev. Lett., 89:258702, (2002).
4. M. Boguñá and R.P.-Satorras. Class of correlated random networks with hidden variables. Phys. Rev. E, 68:036112, (2003).
5. G. Palla, L. Lovász, and T. Vicsek. Multifractal network generator. Proceedings of the National Academy of Sciences, (2010).
6. D. Krioukov, F. Papadopoulos, A. Vahdat, and M. Boguñá. Curvature and temperature of complex networks. Phys. Rev. E, 80:035101, (2009).
7. **S. G. Balogh, P. Pollner and G. Palla: Generalised thresholding of hidden variable network models with scale-free property. Scientific Reports 9, 11273 (2019).**

Constructing large hierarchical networks aiming at realistic, directed and modular structures typical for many kinds of organizations

Fereshteh Rabbani[1], Peter Pollner[2], Gergely Palla[2], and Tamas Vicsek[1,2]

[1] Department of Biological Physics , Eötvös University, H-1117 Budapest, Hungary
[2] MTA-ELTE Statistical and Biological Physics Research Group, H-1117 Budapest, Hungary

1 Introduction

Modular hierarchical structure is a common feature of complex systems within a broad range of ecological, social, communication and economic networks [1–3]. Although many related works have been shown to produce relevant modules in a number of systems [4], the mechanisms leading to the flow hierarchical structure of the modules is still an open problem. Here we propose a new method for constructing modular and directed hierarchical networks based on a successive clustering method. Our approach is based on i) clustering the ”agents” forming the subsequently generated directed network (so that similar agents become close to each other), ii) the clusters form ”modules” and ”elect” a leader, finally, iii) the levels of hierarchy are determined from the number of steps an agent/leader is ”below” the top leaders.

Our study is motivated and supported by observations. These observations can be based both on downloadable data about the structure of organizations and by our everyday personal considerations when we consider how large universities or companies are organized. First of all, in a network representation, they are hierarchical (have ”levels”) and they have mostly two types of links: directed ones from an upper layer to the layer below and ones within a given layer without direction. Here direction stands for a leader-follower relation: the ”bosses” can give instructions to the group of their subordinates. An important observation is that the typical value of organizational units is a number being within 3 to 12 (with an overall average close to 7) [5].

2 Methods

We use clustering as a method of obtaining a multi-level modular structure. Our approach utilizes only a few parameters that allow fine-tuning the size of modules and the number of levels in the hierarchy. Why clustering? Because agents with similar interests, capabilities tend to form groups and our clustering method expresses this fact by bringing together into clusters those units which have close values of their abilities. For example, in a department of material science or a division of car design people with similar knowledge of the related activity form a unit in the organization (within a university, or a car factory, respectively). The whole organization is then made of successively embedded and through these interacting units, and the feature of directionality (along which the communication is exercised) is essential.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2.1 Clustering algorithm

As argued above, a module is made up of people with similar abilities who are committed towards a common purpose. Organizations create modules by grouping individuals in a way that generates a variety of expertise and addresses a specific operational component of the organization. It is important to note that a single network module at a given level itself must be less hierarchical, be almost fully connected and share the leadership. To this end, we put forward a clustering algorithm based on the well known H-K (the Hegselmann-Krause) method [6] bringing gradually closer agents with similar "opinions" and meanwhile resulting in clusters of agents. At first, we represent an organization by a set of N individuals $(i = 1,..,N)$ having a variety of abilities/opinions a taking values between 0 and 1 [7]. The abilities are changing in time according to interaction and influence from neighbors, here is updating rule:

$$a_i(t+1) = \frac{1}{|N_i(t)|} \sum_{j \in N_i(t)} a_j(t) + \eta_i(t) \tag{1}$$

where $N_i(t) = j : |a_j(t) - a_i(t)| \leq \varepsilon_i$ is the neighbor set of agent i at time t and η denotes the level of the added white noise. The summation is over the individuals j whose abilities differ from its own not more than the certain level ε_i. The system uses a similarity measure between nodes to group them onto modules and converges within a number of steps. Besides, by considering a bounded log-normal distribution for ε we could get modules with varying sizes. The updates were implemented synchronously.

2.2 Leader-follower structure

Since the internal structure of an organization can be represented by leader-follower relationships we define a leader for each module after the convergence to a set of clusters occurs. A leader is associated with each module/cluster in a given level as follows: the leader's ability has a value (of the ability of an agent in the cluster) closest to the average ability values in a given module. The clustering is carried out in steps, i.e., the newly "elected" leaders will be clustered in the next stage (example: in a university there are departments, the departments form an institute the institutes are units of a faculty - of, e.g., sciences - the faculties (or "schools") are the main units of the university itself)

2.3 Generating directed, hierarchical modular networks from clustered data

To construct the desired network, we use a simple connection probability function $P_{i,j}$, which is computed from the clustering results through our model (Eq.2). The network is obtained by starting with a set of N nodes already clustered and M adding edges between them in a probabilistic fashion. At the beginning, we assume that there are modules in each level of the network and that the nodes are assigned to a module C_i and level L_i. The connection probability function is formulated as,

$$P_{i,j} = \delta(C_i, C_j)[\ \delta(|L_i - L_j|, 1) + \delta(|L_i - L_j|, 0)] \ + \frac{1}{|L_i - L_j| + B * (|a_i - a_j|^{\alpha})} \tag{2}$$

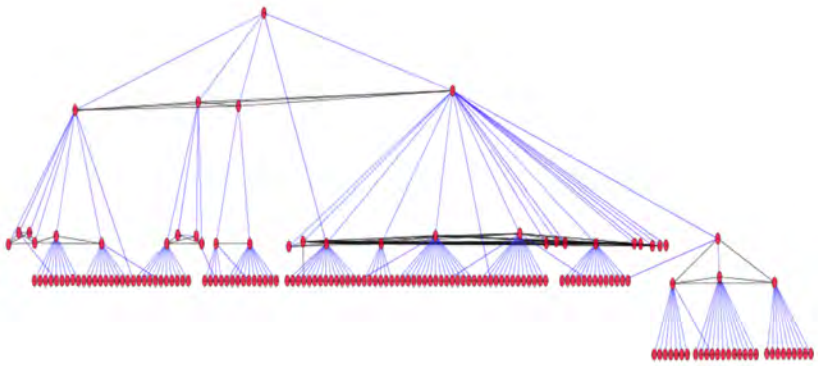


Fig. 1. The directed, modular hierarchical structure for a network with $N = 168$ and $M = 238$ using our connection probability function. This figure is generated by using the following values $B = 10^4$ and $\alpha = 1$. The edges are all directed and top to down. Our approach allows to construct much larger networks of similar structures.

where B and α are constant values and δ denotes the Kronecker function. The network is generated in two steps. First the agents are connected with a probability $P_{i,j}$. This probability function consists of two terms. The first one aims to enforce the agents inside a module to be connected with each other including the leader as well. The second term minimizes the number of connections between distant levels and between agents with too different abilities. In the second step, we update our system to be more realistic. We implement restrictions on the size of the modules and the corresponding levels. We apply specific - not detailed here - measures to maintain realistic cluster sizes and levels. As shown in Fig.1, our method can model a directed hierarchical modular network, with features similar to those which are typical for large organizations.

References

1. Ravasz, E., and Barabási, A. L.: Hierarchical organization in complex networks. Physical review E, 67(2), 026112 (2003).
2. Krause, A. E., FrankK. A., Mason, D. M., Ulanowicz, R. U., and Taylor, W. W.: Compartments revealed in food-web structure. Nature 426, 282 (2003).
3. Ozogány, K. and Vicsek, T.: Modeling the emergence of modular leadership hierarchy during the collective motion of herds made of harems. Journal of Statistical Physics, 158(3), 628-646 (2015).
4. Newman, M.E.: Communities, modules and large-scale structure in networks. Nature physics, 8(1),25 (2012).
5. Fay, N., Garrod, S., and Carletta, J.: Group discussion as interactive dialogue or as serial monologue: The influence of group size. Psychological science, 11(6), 481-486 (2000).
6. Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence models, analysis, and simulation. Journal of artificial societies and social simulation, 5(3) (2002).
7. Zamani, M. and Vicsek, T.: Glassy nature of hierarchical organizations. Scientific reports, 7(1), 1382 (2017).

A random model that relies on maximal bicliques to preserve the overlaps in bipartite networks

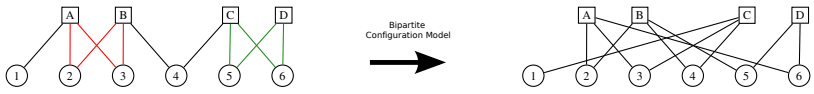
Fabien Tarissan[1]* and Lionel Tabourier[2]

[1] Université Paris-Saclay, CNRS, ENS Paris-Saclay, UMR 7220, ISP, France
[2] Sorbonne Universités, CNRS, LIP6, F-75005 Paris, France

Context. Many real-networks, also refered to as *complex networks*, lend themselves to the use of graphs in order to analyse their structure and model their properties. Since the seminal papers of Barabási and Watts, one usually considers that, whatever the context in which they emerge, all networks share non trivial properties such as a low density, a low average distance, an heterogeneous degree distribution, a high local density, etc.

Such properties distinguish those networks from classic random graph models such as the ones generated by the Erdős-Rényi model which only reproduce the density of the networks. As a consequence, significant effort is dedicated to the elaboration of random models able to capture more intricate properties. Among them, one can cite the Barabási-Albert model which succeeds in producing a heterogeneous (scale-free) degree distribution but fail in generating graphs with a high local density, the Watts and Strogatz model which generates networks with the opposite features or the Configuration Model [3] which generate random graphs with a prescribed degree sequence but with a low local density. All in all, and despite the different attempts, generating a graph exhibiting all expected properties is still an open issue.

The purpose of this study is to present a new step toward that goal by exploiting the bipartite version of the configuration model. Indeed, although useful, the representation of networks as unipartite graphs does not account for the inherent complexity induced by the hierachical structure observed in most real networks. This observation led the scientific community to turn to *bipartite graphs* to describe such complex structure when possible. This formalism allows to define explicitly two disjoint sets of nodes and the links only relate a node of one set to a node of the other set. The natural extension of the configuration model to bipartite graphs allows to preserve the degree of every nodes while shuffling the links, as depicted below:



However, as illustrated in the picture, such a model can easily disturb key patterns of the structure. Although the degree distribution is preserved, the two bicliques (in red and green) completely vanish after the randomization due to a slight modification of

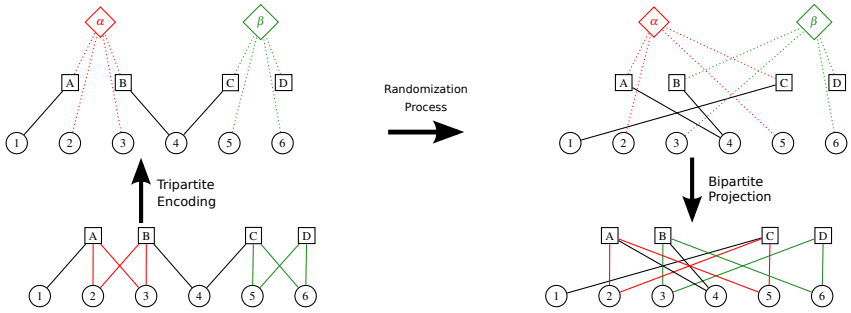
*This work is funded in part by CNRS under grant n 245 709 (PICS project Récital).

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the links. To that regard, recent studies showed that overlaps (top nodes connected to common bottom nodes) are ubiquitous and important patterns in bipartite networks [4].

In order to overcome this issue, we propose in this paper a generative model able to preserve both the degree sequence and the overlaps of real networks. It relies on the encoding of those patterns in a third level, defining thus a *tripartite* graph, on which we perform the randomization. More precisely, we first perform the enumeration of all maximal bicliques in the bipartite graph, then encode the bicliques in a third level before performing a randomization preserving the encoding. Finally, we project the obtained tripartite graph into its corresponding bipartite structure:



One key operation in this method relies on the tripartite encodings of the bipartite structure. We tested several natural heuristics which select the bicliques in a given order to create the tripartite encoding: a random selection, a selection that maximizes the number of links encoded and one that maximizes the number of nodes captured.

Results show that all heuristics lead to generating bipartite graphs in which the overlaps are preserved. We show in addition that several other properties emerge naturally with much more accuracy than with a standard bipartite configuration model.

Results. In order to validate the approach, we tested the models on 9 datasets that have an underlying bipartite structure. Due to space limitation, we only show the results on three representative datasets: `HepB` is a network featuring scientists and the articles that they coauthored, collected from Medline repository using the keyword *Hepatitis B*, `BPSE` is a network built from the proteins of bacteria *Burkholderia pseudomallei* and the biochemical reactions they take part in, and `Youtube` contains the membership of Youtube users as collected in 2007 [2].

For each network, we computed several properties both on the original bipartite graphs and on the ones generated by the models. More precisely, let $G = (\top, \bot, E)$ be a bipartite graph, where $\top$ is the set of *top* nodes, $\bot$ the set of *bottom* nodes, and $E \subseteq \top \times \bot$ the set of links between $\top$ and $\bot$. We denote by $N(u)$ the set of neighbors of u in the bipartite graph and by $N_2(u)$ its neighbors at distance 2. We computed several nodes characteristics related to the overlaps: the *bipartite coefficient* [1] based on the Jaccard index defined as $\texttt{bip}(u) = \frac{\sum_{v \in N(u)} \texttt{cc}(u,v)}{|N(u)|}$ where $\texttt{cc}(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$, the *dispersion*

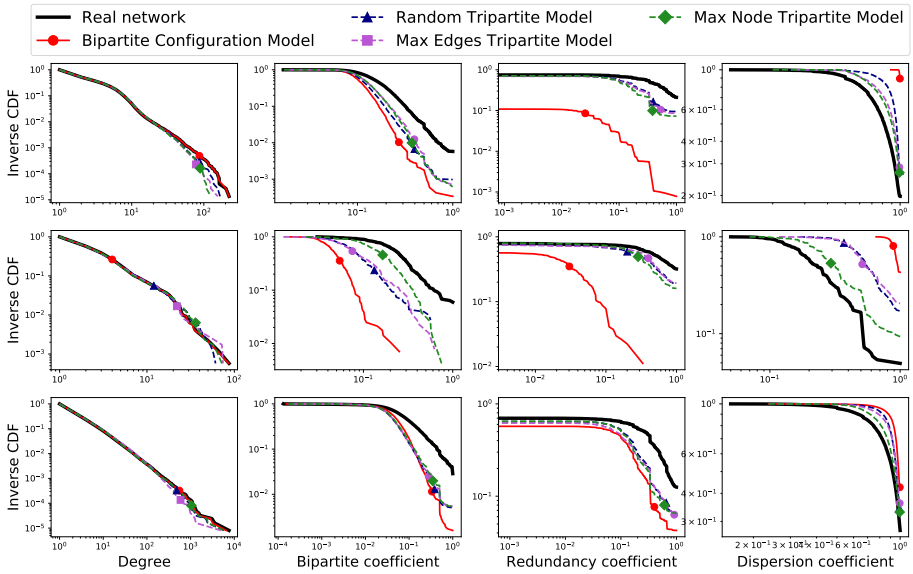


Fig. 1. Inverse cumulative distribution of the degree distribution (first column), the bipartite clustering coefficient (second column), the redundancy coefficient (third column) and the dispersion coefficient (fourth column) for `HepB` (top), `BPSE` (middle) and `Youtube` (bottom).

coefficient [4] defined as $\texttt{disp}(u) = \frac{|N_2(u)|}{\sum_{v \in N(u)} (|N(v)|-1)}$ and the *redundancy coefficient* [1] defined as $\texttt{rd}(u) = \frac{|\{(v,w) \in N(u) \times N(u) \text{ s.t. } \exists u' \neq u, (u',v) \in E \text{ and } (u',w) \in E\}|}{\frac{|N(u)|(|N(u)|-1)}{2}}$.

Figure 1 presents the results for the 3 datasets and the distribution of all characteristics considered. For all features examined here the tripartite models succeed in preserving the properties better than the configuration model applied on the bipartite structure. This is particularly true for the redundancy and dispersion coefficients.

References

1. M. Latapy, C. Magnien, and N. Del Vecchio. Basic notions for the analysis of large two-mode networks. Social Networks, 30(1):31–48, January 2008.
2. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07), San Diego, CA, October 2007.
3. M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graphs with arbitrary degree distribution and their applications. Phys. Rev. E., 64, July 2001.
4. R. Tackx, F. Tarissan, and J.-L. Guillaume. Revealing intricate properties of communities in the bipartite structure of online social networks. In Proceedings of the 2015 IEEE 9th International Conference on Research Challenges in Information Science, RCIS'15, pages 321–326. IEEE, 2015.

Nonlinear interactions in noisy coevolving networks

Tomasz Raducha[1,2] and Maxi San Miguel[2]

[1] Institute of Experimental Physics, Faculty of Physics, University of Warsaw, Pasteura 5, 02-093 Warsaw, Poland, raducha.tomasz@fuw.edu.pl,
[2] IFISC, Instituto de Física Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat Illes Balears, E-07122, Palma de Mallorca, Spain

1 Introduction

Coevolving or adaptive network models are increasingly popular due to their closer relation with real-world systems in comparison with static or evolving networks [1]. Most of the empirical networks display both the topological evolution and state's dynamics. Moreover, a nontrivial feedback loop between these aspects renders a simple sum of effects analyzed separately incomplete. Adaptive mechanisms can generate results unreachable when omitting one side of the system. The question is how important are the new outcomes, as there is always a trade-off between the model's complexity and effectiveness. Recent works have proved that this balance favours coevolving models [1]. They not only have the microscopic assumptions in better agreement with empirical observations, but also produce macroscopic results otherwise unachievable.

Here we aim at exploring the joint effect of the three important aspects of real-world networks – the coevolution of structure and state, the non-linearity of interactions and the noise – on the behavior of the system. As the framework we choose a simple yet fruitful in explaining empirical observations model, namely the voter model [2]. With a binary state, it provides a convenient platform for analysis of opinion dynamics. Exceptionally, fairly distinct phenomena like ants behavior [3], or stock market nature [4] were successfully described within the frames of the voter model. It has been already extended for coevloution, noise and non-linearity separately [2, 5, 6]. The joint effect of these aspects, however, can be more complex than a superposition of so-far obtained results. Hereafter we seek to examine it.

2 The model

First a random graph is generated and every node is assigned a state $s_i \in \{-1, +1\}$. In every time step a node is chosen at random, we call it the active or focal node. Then, with probability $(a_i/k_i)^q$ an interaction occurs, where k_i is the degree of the focal node i, a_i is the number of neighbours of the node i being in the opposite state, and q is the non-linearity parameter of the model. If an interaction occurs, one of the a_i neighbors in a different state is chosen, call it j. Then, with probability p a link rewiring is performed and with complementary probability $1-p$ a state copying. The link rewiring is global and random. At the and of the time step, regardless of what happened before, the active node with probability ε draws a random state. The algorithm of the model is illustrated in the Figure 1.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

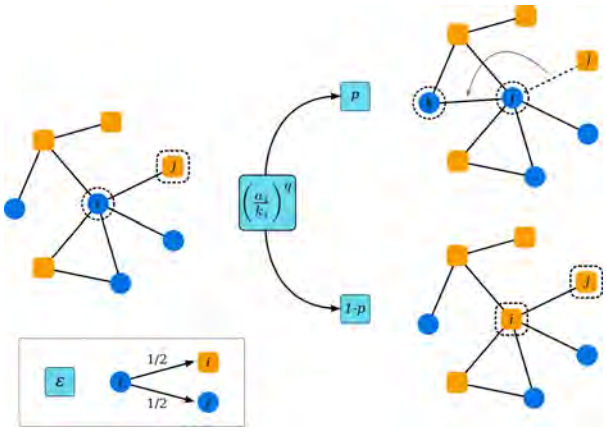


Fig. 1. Schematic illustration of update rules in the nonlinear coevolving voter model with noise.

3 Results

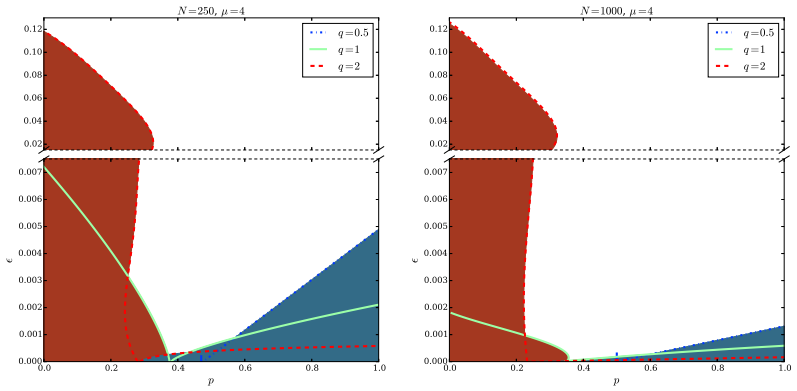


Fig. 2. Phase diagram in p-ε space for $N = 250$ (left) and $N = 1000$ (right) for $\mu = 4$ and different values of q. Picture is made based on simulations averaged over 500 realizations.

We numerically study the p-ε phase diagram for three different values of the q parameter – the sub-linear case $q = 0.5$, the ordinary linear case $q = 1$, and the super-linear case $q = 2$. These phase diagrams are presented in the Figure 2 for two different network sizes. We can distinguish three general phases in the model. The phase A, indicated by the red area in the figure, is a consensus phase. In this range of parameters the network stays in a consensus state for most of the time, i.e. magnetization is close to ± 1 and the network is connected having one large component. Obviously, for any finite amount of noise in the system a frozen configuration does not exist and a phase is described by its dynamical stationary state. If we increase the noise rate ε or the plasticity p sufficiently, we obtain the fully-mixing phase B, indicated by the white

area in the Figure 2. Here, the magnetization drops to zero $m = 0$, hence there is no consensus in the system anymore. But the network still stays connected most of the time. Finally, for high values of the rewiring probability above p_c and not too big noise rates the phase C arises. It is marked by the blue area in the figure. In this region we report a dynamical fragmentation – the network consists of two separate components being in the opposite states. It is possible, however, that two components get connected for a moment due to the noise and random rewiring, creating again one big network. The phase C can be described as dynamical switching between these two arrangements.

We derive equations (1) governing the dynamics of the system and describing time evolution of the magnetization m and the interface density ρ. Several solutions (m^*, ρ^*) can be found depending on the parameter choice, however not all of them are stable, therefore not all of them are observed in experiments. Since the analytical description is derived for the thermodynamic limit, we don't observe stable fixed points at non-zero magnetization for $q \leq 1$. This finding is consistent with the scaling behavior of the numerical results indicating existence only of the phase B in the large network limit.

$$\begin{aligned}
\frac{dm}{dt} &= 2(1-p)(1-\varepsilon)(n_- n_q^- - n_+ n_q^+) + \varepsilon(n_- - n_+), \\
\frac{d\rho}{dt} &= \frac{2}{\mu}\left[(1-p)(1-\varepsilon)(n_+ n_q^+ \delta_+ + n_- n_q^- \delta_-) - p(n_+ n_q^+ + n_- n_q^-) + \frac{\varepsilon}{2}(n_+\delta_+ + n_-\delta_-)\right].
\end{aligned} \tag{1}$$

Our work fills the gap in the studies of the CVM. It provides a binding between studies of the CVM with noise [5] and studies on the nonlinear CVM [6]. Additionally, it collapses to the nonlinear noisy voter model [7] and the ordinary CVM [2] for a proper configuration of parameters values. We obtain full consistency with those limit cases and explore untouched regions in between. Our work brings the analysis of the voter model to a greater complexity by taking into account many possible effects. It may provide a tool in evaluation of the relevance of different factors in description of opinion dynamics, but can be also a reference point in the study of coevolving network models.

References

1. Gross, T., Blasius, B.: Adaptive coevolutionary networks: a review. Journal of the Royal Society Interface **5**(20) (2008) 259–271
2. Vazquez, F., Eguíluz, V.M., San Miguel, M.: Generic absorbing transition in coevolution dynamics. Physical review letters **100**(10) (2008) 108702
3. Kirman, A.: Ants, rationality, and recruitment. The Quarterly Journal of Economics **108**(1) (1993) 137–156
4. Alfarano, S., Lux, T., Wagner, F.: Estimation of agent-based models: the case of an asymmetric herding model. Computational Economics **26**(1) (2005) 19–49
5. Diakonova, M., Eguíluz, V.M., San Miguel, M.: Noise in coevolving networks. Physical Review E **92**(3) (2015) 032803
6. Min, B., San Miguel, M.: Fragmentation transitions in a coevolving nonlinear voter model. Scientific Reports **7**(1) (2017) 12864
7. Peralta, A.F., Carro, A., San Miguel, M., Toral, R.: Analytical and numerical study of the non-linear noisy voter model on complex networks. Chaos: An Interdisciplinary Journal of Nonlinear Science **28**(7) (2018) 075516

The role of driving signal in the evolution of social networks

Ana Vranić and Marija Mitrović Dankulov

Scientific Computing Laboratory, Center for Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia

1 Introduction

Dynamics and emergence of collective behavior in social system strongly depend on the structure of the interactions between actors in the social network. The nature of connections has been studied through empirical analysis and theoretical models of evolving networks [1]. Models of evolving networks start with one, or the small number of randomly connected nodes. The network grows by the addition of new nodes, which link to ones already present in the system, following some linking rule. These rules can shape the network in a specific way. For example, the preferential attachment mechanism is essential for reproducing the networks with a heterogeneous distribution of the number of first neighbors.

The role of driving, i.e., non-constant addition of new nodes in the network is still poorly understood. While standard networks models assume that the addition of new nodes is constant in time, the growth signal of real social systems varies and influences network structure [2]. It is of great importance to understand the interplay between the driving signal and network topology, and how they, separately and in combination, shape the collective behaviour in social systems. We use a model of network with aging nodes to examine the role of driving signal in a network.

2 Results

The aging model incorporates the time in a non-trivial manner by introducing nodes aging [3]. The network is generated by adding one node with one link to the target node in each time step, t. Probability for connecting new node in the network depends on degree k of the target node and the age difference τ between the new and target node,

$$\Pi_i(t) \sim k_i(t)^{\beta} \tau_i^{\alpha} \tag{1}$$

Different values of parameters α and β lead to networks with different structural properties.

We customised the aging model by allowing the addition of multiple nodes $(M > 1)$ and links $(L > 1)$, in each time step. As input in the simulation, we used the driving signal from the Meetup website, TECH social group [4]. Driving signal shows the number of new members that joined a group at a single event.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

We run the simulations for TECH signal and randomized TECH signal, for all combinations of parameters $-3 < \alpha < 0$ and $1 < \beta < 3$, generating a sample of 100 networks. New members in network can make one ($L = 1$), or more ($L = 3$) connections. As the average number of added nodes per time step is $M = 1$, we looked into differences of networks driven with original and randomized TECH signal and ones with constant growth in the time. We use dissimilarity measure (D-distance) [5] to compare samples of networks grown with different signals. D-distance considers Jensen-Shenon divergence and node distance distribution.

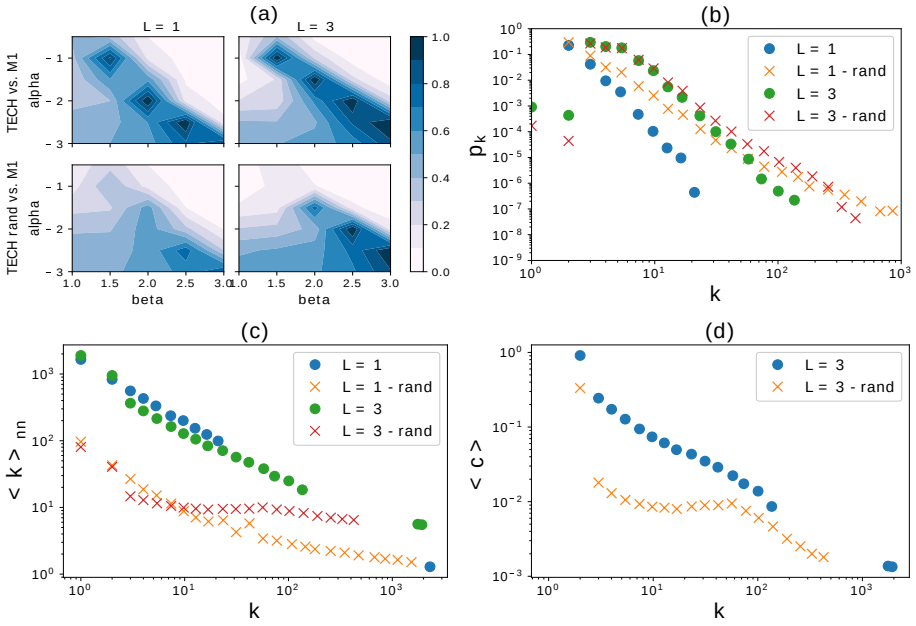


Fig. 1. (a) Dissimilarity distance between networks with (randomized) TECH and constant $M = 1$ signal, for number of links $L = 1$ and $L = 3$ in α-β plain. Network properties of (randomized) TECH signal for different values of L: (b) degree distribution, (c) dependence of average neighbor degree on node degree, (d) node clustering coefficient; for fixed model parameters $\alpha = -1.$ and $\beta = 1.5$

Figure **1(a)** shows calculated D-distance between networks obtained for original driving signal vs. $M = 1$ (upper panel) and for randomized driving signal vs. $M = 1$ (lower panel). We notice a critical region around $\beta = 1.5$ and $\alpha = -1$, where D-distance between TECH and M1 signal is greater than between randomized TECH signal and M1. For these parameters, we represent the topological features of networks. For degree distribution (**Fig.1(b)**) we observe the only difference in slope between original and randomized TECH signal, with linking parameters $L = 1$ and $L = 3$. Networks generated with the original and reshuffled signal have significantly different topology if we compare degree-degree correlations and clustering coefficient.

Networks obtained for the real signals are strongly disassortative (**Fig.1(c)**) and have hierarchical structure, i.e., their clustering coefficient (**Fig.1(d)**) decreases with k. On the other hand, networks observed from driving the model with the randomized signal are uncorrelated, and their clustering weakly depends on the degree. Networks generated with the aging network model for L=1 are tree-like networks. They don't have triangles and their clustering is equal to 0.

Summary. Our results show that for the certain values of model parameters networks obtained from the driving with original signals have different topological features than ones obtained from the driving with random signals, although they evolve under the same linking rules. We find that driving signals alter the shape of the degree distribution, degree-degree correlations and clustering in the network. The effect is the largest for the values of model parameters for which we obtained networks with broad degree distribution. This difference disappears as we move away from these parameters. Our results strongly support the conclusion that driving signal is an important factor in the evolution of social networks and it has to be included, as a parameter, in modeling social systems.

References

1. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU, Complex networks: Structure and dynamics, Phys. Rep. 424, 175-308 (2006).
2. Mitrović Dankulov M, Melnik V, Tadić B:The dynamics of meaningful social interactions and the emergence of collective knowledge. Scientific Reports volume 5, Article number: 12197 (2015)
3. Basu Hajra K, Sen P.:Phase transitions in an aging network. Phys. Rev. E 70, 056103 (2004)
4. Smiljanić J, Mitrović Dankulov M: Associative nature of event participation dynamics: A network theory approach, PLoS ONE 12, e0171565 (2017)
5. Schieber et. all (2017).: Quantification of network structural dissimilarities. Nature Communications. 8. 13928. 10.1038/ncomms1392 (2017)

Long-range degree correlations of fractal clusters in random networks

Shogo Mizutaka and Takehisa Hasegawa

Department of Mathematics and Informatics, Ibaraki University,
Bunkyo 2-1-1 Mito, Japan
shogo.mizutaka.sci@vc.ibaraki.ac.jp
takehisa.hasegawa.sci@vc.ibaraki.ac.jp

1 Introduction

Networks consisting of nodes and edges are observed in nature and society. From the viewpoint of the relation between the average path length and the system size (the total number of nodes in a network), real-world networks can be classified into small-world and fractal networks [1]. In small-world networks, the average path length increases with the logarithm of the system size at most. In fractal networks, the number of boxes required to cover a network decreases with the radius of boxes in a power law fashion.

As has been reported, fractal networks have a long-range degree correlation that is a degree correlation between two nodes not directly connected; fractal networks have long-range anti-correlations in the sense of degree fluctuation [2]; nearest neighbor degree correlations fail to explain the fractality of a network [3]. However, it is not sufficient to understand the degree-correlated structure of fractal networks because it is difficult to handle the long-range degree correlation both analytically and numerically in most networks.

We treat the long-range degree correlations of an infinitely large cluster extracted from the Erdős-Rényi random graph. In the Erdős-Rényi random graph with the degree distribution $p_k = \bar{k}^k e^{-\bar{k}}/k!$, where $\bar{k}$ is the average degree of the graph, there exists (does not exist) the infinitely large cluster if the average degree $\bar{k}$ is greater (less) than the critical value $\bar{k}_c = 1$. At the critical average degree $\bar{k}_c$, the cluster becomes fractal and such a cluster is hereinafter referred to as the fractal cluster. Using the generating functions, we obtain the property of the fractal cluster extracted from the critical Erdős-Rényi random graph. Specifically, we derive the probability $P_{fc}(k,k'|l)$ that two randomly chosen nodes separated by distance l from each other have the degrees k and k' on the fractal cluster and characterize the long-range degree correlations of the fractal cluster extracted from the Erdős-Rényi random graph.

2 Results

Let us consider an infinite Erdős-Rényi random graph which is degree-uncorrelated and is locally tree-like. We introduce the probability u that an edge does not lead an infinitely large cluster. In the Erdős-Rényi random graph, the probability u is given as $u = \sum_k k p_k u^{k-1}/\bar{k} (= \sum_k p_k u^k)$. From the probability u, we have the probability $S = 1 -$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

$\sum_k p_k u^k = 1-u$ that a node belongs to an infinitely large cluster. For $\bar{k} < \bar{k}_c$, the network consists of only small (finite) clusters, which corresponds with $u = 1$. For $\bar{k} > \bar{k}_c$, there exists an infinitely large cluster. In this case, the relation $u < 1$ is satisfied. Using the probability u, we can extract properties of the infinitely large cluster, e.g., the relative size S of the cluster [4], the degree distribution and the joint probability that degrees of two ends of a randomly chosen edge are k and k' [5, 6]. We derive the probability that two randomly chosen nodes separated by distance l from each other have the degrees k and k' on the infinitely large cluster. Approaching the system to the critical average degree, i.e., $\bar{k} \to \bar{k}_c$, the probability $P_{\rm fc}(k,k'|l)$ on the fractal cluster behaves as

$$P_{\rm fc}(k,k'|l) = \frac{\bar{k}_c(l-1)+(k+k'-2)}{\bar{k}_c(l-1)+2} \frac{kp_k}{\bar{k}_c} \frac{k'p_{k'}}{\bar{k}_c}. \tag{1}$$

Figure 1 shows $P_{\rm fc}(k,k'|l)$ for the critical Erdős-Rényi random graph as a function of k and k' for several distances. Wireframes (analytical treatment (1)) match perfectly with symbols (simulation results), which implies the validity of our analytical treatment.

Fig. 1. Probability distribution $P_{\rm fc}(k,k'|l)$ for the critical Erdős-Rényi random graph ($\bar{k} = \bar{k}_c$) as a function of k and k' for $l = 1$, 3, and 5. Wireframes are results for analytical calculations obtained by Eq. (1) and symbols represent corresponding simulation results. In simulations, the number of nodes is set as 10^7 for a single sample.

Summary. We discuss the long-range degree correlations of the fractal cluster in the critical Erdős-Rényi random graph. From Eq. (1), we can obtain the average degree $k_l^{\rm fc}(k)$ of l distant nodes from a degree-k node on the fractal cluster which is a generalization of the average nearest neighbor degree $k_{\rm nn}(k)$ of nodes with degree k. The behavior of $k_l^{\rm fc}(k)$ shows that the fractal cluster possesses a negative degree correlation for any distance l. In this presentation, we will present the general result for random graphs with arbitrary degree distributions and discuss long-range degree correlations of not only the fractal cluster but also the infinitely large cluster.

Acknowledgement

Authors acknowledge financial support from JSPS (Japan) KAKENHI Grant Number JP18KT0059. S.M. acknowledges supported by a Grant-in-Aid for Early-Career Scientists (No. 18K13473) and a Grant-in-Aid for JSPS Research Fellow (No. 18J00527)

from JSPS (Japan). T.H. acknowledges financial support from JSPS (Japan) KAKENHI Grant Number JP19K03648.

References

1. Cohen, R., Havlin, S.: Complex networks: structure, robustness and function. Cambridge university press (2010)
2. Rybski, D., Rozenfeld, H.D., Kropp, J.P.: Quantifying long-range correlations in complex networks beyond nearest neighbors. Europhys. Lett. 90 (2), 28002 (2010)
3. Fujiki, Y., Mizutaka, S., Yakubo, K.: Fractality and degree correlations in scale-free networks. Eur. Phys. J. B 90, 126 (2017)
4. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 64, 026118 (2001)
5. Bialas, P., Oleś, A.K.: Correlations in connected random graphs. Phys. Rev. E, 77, 036124 (2008)
6. Tishby, I., Biham, O., Katzav, E., Kűhn, R.: Revealing the Micro-Structure of the Giant Component in Random Graph Ensembles. Phys. Rev. E 97, 042318 (2018)

Reconstructing the history of growing trees

Gábor Timár[1], Rui A. da Costa[1], Sergey N. Dorogovtsev[1,2], and José F. F. Mendes[1]

[1] University of Aveiro, Aveiro, Portugal
[2] A. F. Ioffe Physico-Technical Institute, Russia
gtimar@ua.pt

1 Introduction

Models of network evolution have been studied intensively in the past two decades, and have resulted in an increasingly thorough understanding of the structure and function of various classes of real-world networks, both natural and artificial. Most theoretical works on evolution mechanisms have followed a deductive approach, focusing on structures resulting from general network evolution rules. The recent surge of scientific activity related to artificial intelligence and machine learning algorithms has led to the new field of *network archeology*, where the aim is to infer information about the history of a network from a current static snapshot of its structure. Such information can help us better understand its current structure and predict future states.

Inferring the history of growing networks, even considering the simplest growth mechanisms and network structures, is a difficult combinatorial problem. In the last decade considerable literature has accumulated around the problem of inferring the root of a growing tree, one of the simplest well-defined problems in network archeology. An efficient root detection method was introduced in [1] and was shown to give the exact maximum likelihood estimate of the root of a randomly growing tree confined to a Bethe lattice of arbitrary coordination number. This method has also been shown to be exact for trees grown according to the Barabási-Albert model [2], and its accuracy has been studied in various non-exact scenarios [5–7]. A far more complex problem is the inference of the complete history of a growing network. Very recently some principled methods have been suggested, based on a Bayesian inference framework [3, 4]. These approaches rely on Monte Carlo sampling from an appropriately weighted distribution of possible histories, which are computationally demanding and hence not easily scalable. Simpler heuristic approaches have also been suggested in [3, 4], based on node degrees and other simple centrality measures that are efficiently retrieved from network structure.

2 Results

2.1 Root inference

We consider a root finding algorithm based on the concept of history degeneracy: the number of distinct node sequences (complete node orders) that start at a given node and produce exactly the tree under observation. The structure of the given tree imposes a partial order on the set of its nodes, and there are a large number of particular sequences

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

that comply with that partial order. (Such a complete sequence is also called a linear extension of the given partially ordered set (poset)). The probability P_i that a given node i of a tree was the root is proportional to the number $\mathscr{N}_i$ of allowed sequences generating our given tree, started at node i. We derive a set of message passing equations whose solutions $Q_{i\leftarrow j}$ combine to give the above probabilities on an arbitrary tree,

$$P_i \sim \mathscr{N}_i \sim \prod_{j\in\partial i} Q_{i\leftarrow j}, \tag{1}$$

where ∂i denotes the set of neighbours of node i. The solutions $Q_{i\leftarrow j}$ can be written in vector form as

$$\ln \mathbf{Q} = (\mathbf{B}-\mathbf{I})^{-1} \ln \mathbf{N}, \tag{2}$$

where $\ln \mathbf{Q}$ is the element-wise logarithm of the vector $\mathbf{Q} = \{Q_{i_1\leftarrow j_1}, Q_{i_2\leftarrow j_2}, \ldots\}$. Similarly $\ln \mathbf{N} = \{\ln N_{i_1\leftarrow j_1}, \ln N_{i_2\leftarrow j_2}, \ldots\}$, where $N_{i\leftarrow j}$ denotes the number of nodes in the branch "upstream" of directed link $i \leftarrow j$. $\mathbf{B}$ denotes the nonbacktracking matrix of the given tree and $\mathbf{I}$ is the identity matrix. The probabilities calculated in this way are proportional to the rumor centrality of [1]. We show that this method is exact for general linear preferential attachment (LPA) trees.

This message passing scheme also allows us to accurately measure the history degeneracy of LPA trees and random trees grown in a Bethe lattice. We call the latter constrained random recursive (CRR) trees. In both classes the logarithm of the history degeneracy $\ln \Omega$ behaves as

$$\ln \Omega \cong N \ln N - aN, \tag{3}$$

where N is the tree size and a is a size-independent constant. The behaviour of a as a function of the parameter of CRR and LPA trees is shown in Fig. 1.

2.2 Reconstruction of complete history

Based on the above scheme we propose a fast algorithm to reconstruct the complete history of a growing tree. Our method is a step-wise maximum likelihood estimate of the history that is calculated in each step according to a slight modification of Eq. (2), taking into account the part of the tree that already exists (is already inferred), and nodes whose order is yet to be determined. Our method works well for trees with low history degeneracy, and the reconstruction quality gets progressively worse as degeneracy increases, see Fig. 1.

For low degeneracy trees our method works considerably better than simple degree-based reconstruction methods as indicated by comparison of the average relative overlap of inferred sequences with the real one, and by measuring the correlation of inferred ranks with the real ranking.

We expressed the probabilities of nodes on a growing tree being the root using the nonbacktracking matrix of the given tree. We showed that this is exactly the maximum likelihood estimate in the case of general linear preferential attachment trees. We numerically studied the efficiency of this root finding approach and proposed a fast method

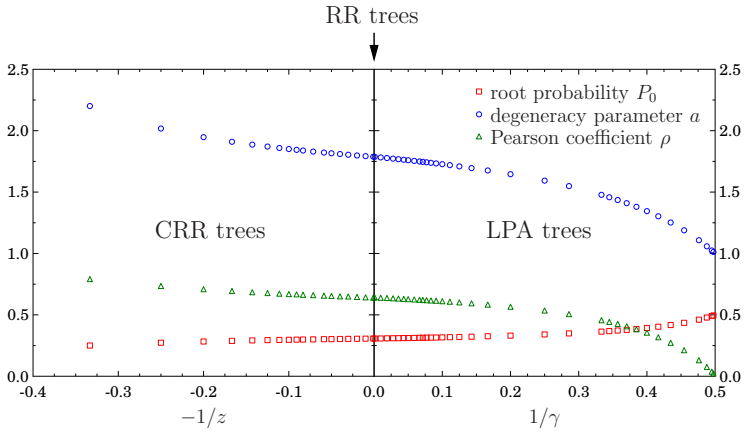


Fig. 1. Probability P_0 of being the root for the most likely node, the history degeneracy parameter a (see Eq. (3)) and Pearson rank correlation ρ, for the two classes of trees considered: CRR trees parametrized by the coordination number z of the underlying Bethe lattice, and LPA trees parametrized by the degree distribution exponent γ. The detectability of the root node monotonically increases, and the reconstructability of the complete history monotonically decreases as we approach a star structure ($1/\gamma = 0.5$). Note that the limits $z \to \infty$ and $\gamma \to \infty$ both correspond to random recursive (RR) trees.

to infer the complete history of a growing tree, based on the same principle of counting history degeneracies. We accurately measured the history degeneracies of linear preferential attachment trees and random trees grown in Bethe lattices, and concluded that high-quality history reconstruction is possible in the low-degeneracy range.

References

1. Shah, M., Zaman, T.: Rumors in a Network: Who's the culprit? IEEE T. Inform. Theory 57(8), 5163–5181 (2010)
2. Bubeck, S., Devroye, L., Lugosi, G.: Finding Adam in random growing trees. Random Struct. Algor. 50(2), 158–172 (2017)
3. Young, J.-G. et al.: Network archeology: Phase transition in the recoverability of network history. arXiv:1803.09191 (2019)
4. Sreedharan, J. K. et al.: Inferring Temporal Information from a Snapshot of a Dynamic Network. Sci. Rep. 9(1), 3057 (2019)
5. Luo, W., Tay, W. P., Leng M.: Identifying Infection Sources and Regions in Large Networks. IEEE Trans. Signal Process. 61(11), 2850–2865 (2013)
6. Dong, W., Zhang, W., Tan, C. W.: Rooting out the Rumor Culprit from Suspects. IEEE Int. Symp. Inf. Theory 2671 (2013)
7. Shah, M., Zaman, T.: Finding rumor sources on random trees. Oper. Res. 64(3), 736–755 (2016)

Finding the optimal nets for self-folding Kirigami

N. A. M. Araújo[1,2], R. A. da Costa[3], S. N. Dorogovtsev[3,4], and J. F. F. Mendes[3]

[1] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2] Centro de Física Teórica e Computacional, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[3] Department of Physics & I3N, University of Aveiro, 3810-193 Aveiro, Portugal
[4] A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

1 Introduction

The synthesis of three-dimensional polyhedral shells at the micron and nano scales is key for encapsulation and drug delivery. Inspired by the Japanese art of Kirigami, where hollowed structures are obtained from cutting and folding a sheet of paper, lithographic methods have been developed to form shells from two-dimensional templates of interconnected panels. The potential is enormous, for a wide range of shapes and sizes can be obtained.

Ideally, the unfolded templates (nets) should spontaneously self-fold into the target structure to reduce production costs and achieve large-scale parallel production. However, some nets are much more likely to self-fold into the desired shell under random movements [1]. The optimal nets are the ones that maximize the number of vertex connections, i.e., vertices that have only two of its faces cut away from each other in the net.

Even a shell of moderate size (in number of faces) has many possible nets, but only a small fraction of them is optimal. For example, for the dodecahedron, with only twelve faces, less than 0.04% of its more than 5 million nets are optimal, i.e., to obtain an optimal net one would need to randomly sample 2500 configurations on average.

2 Methods and Results

Previous methods for finding such nets are based on random search and thus do not guarantee the optimal solution. Adapting concepts and methods from Graph Theory, we show in [2] that the optimal solution can be obtained in a deterministic and systematic manner. We map the connectivity of the shell into a shell graph, where the nodes and links of the graph represent the vertices and edges of the shell, respectively, see Fig. 1. Identifying the nets that maximize the number of vertex connections corresponds to finding the set of maximum leaf spanning trees of the shell graph.

As we showed in [2], the fraction of nets that have the maximum number of vertex connections decays exponentially with the number of edges in the polyhedron, reinforcing the necessity of a deterministic method. This method allows not only to design the self-assembly of much larger shell structures but also to apply additional design criteria, as a complete catalogue of the maximum leaf spanning trees is obtained.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

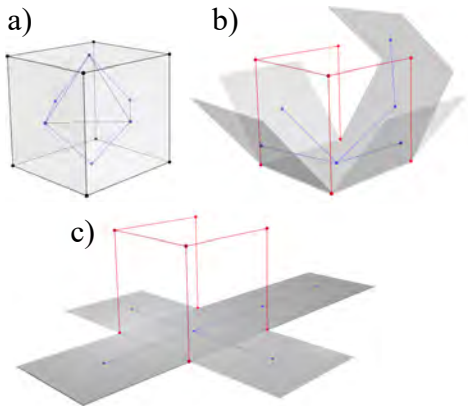


Fig. 1. Net of a cubic shell. (a) The cubic shell is mapped into a *shell graph* (black), where nodes and links of the *shell graph* are the vertices and edges of the polyhedron, respectively. In the *face graph* (blue), the nodes are the shell faces and the links connect pairs of adjacent faces. To unfold the shell into a two-dimensional template (net), one needs to remove a set of shell edges (e.g., red links in (b) and (c)). This set of removed shell edges is a sub-graph of the shell graph. The set of removed shell edges (cut) and the net are spanning trees of the shell and face graphs, respectively. The four vertices of the bottom face are vertex connections.

Using the minimization of the radius of gyration as the secondary design criteria, we find the optimal net for several examples of shells [2], some of which we show in Fig. 2. Moreover, we develop a variation of the method for open shells structures, i.e. with some of the polyhedral faces missing, who's optimal nets are also shown is Fig. 2.

(1)

Three-dimensional shells can be synthesized from the spontaneous self-folding of two-dimensional templates of interconnected panels, called nets. However, some nets are more likely to self-fold into the desired shell under random movements. Previous methods for finding such nets are based on random search. Here, we propose a deterministic procedure. This method allows us not only to design the self-assembly of much larger shell structures, closed and open, but also to apply additional design criteria.

References

1. Pandey, S., Ewing, M., Kunas, A., Nguyen, N., Gracias, D. H., Menon, G.: Algorithmic design of self-folding polyhedra. PNAS, 108(50), 19885–19890 (October 2011).
2. Araújo, N.A.M., da Costa, R.A., Dorogovtsev, S. N., Mendes, J.F.F.: Finding the optimal nets for self-folding kirigami. Phys. Rev. Lett. 120, 188001 (May 2018)

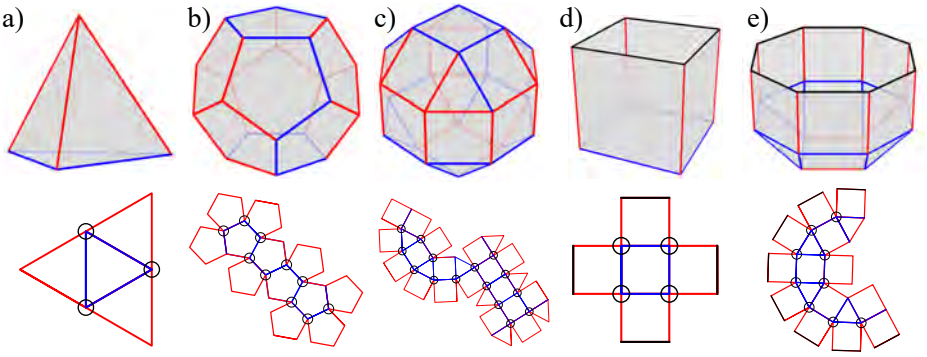


Fig. 2. Five examples of shells and of one of their nets corresponding to a cut that is a maximum leaf spanning tree: a) tetrahedron, with four faces and nine edges, it has one non-isomorphic optimal net; b) dodecahedron, with twelve faces and thirty edges, it has 21 non-isomorphic optimal nets; c) small rhombicuboctahedron, with 26 faces and 48 edges, it has 32 non-isomorphic optimal nets; d) open cubic shell, with five faces and twelve edges, it has only one optimal net; e) small rhombicuboctahedron with the top nine faces removed and 17 faces, 36 edges and 20 nodes remaining, it has 90 non-isomorphic optimal nets. The black circles in the nets indicate the vertex connections.

Distances in Node Duplication networks

Chanania Steinbock, Ofer Biham and Eytan Katzav

Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel,
eytan.katzav@mail.huji.ac.il

To gain insight into the structure of complex networks, it is useful to study their growth dynamics. In general, it appears that many of the networks encountered in biological, ecological and social systems grow step by step, by the addition of new nodes and their attachment to existing nodes. In some networks, the new nodes emerge with no predefined connections, while in other networks the new nodes result from the duplication of existing nodes, followed by a stochastic readjustment of their links.

The effect of node duplication (ND) processes on the structure of complex networks was studied using the ND model [1–5]. In this model, at each time step a random node, referred to as a mother node, is selected for duplication. The new, daughter node, retains a copy of each link of the mother node with probability p. Furthermore, in an important variant referred to as the corded ND model (introduced in Refs. [6, 7]), the daughter node forms a link to the mother node (deterministically) as well. Examples of the resulting network are shown in Fig. 1. An important extension was proposed and studied recently in our work [8], where the links are directed - pointing from the daughter node to the mother node (and only outgoing arcs are copied).

The degree distributions of these network turn out to follow a power-law distribution, with an exponent that depends continuously on p - thus the corded ND networks are scale-free [6–8]. The undirected version is suitable for the study of acquaintance networks, in which a newcomer who has a friend in a new community becomes acquainted with members of the friend's social group [9]. The main advantage is that the formation of triadic closures is built into the dynamics of the undirected model. This means that once the daughter node forms a link to a neighbor of the mother node, it completes a triangle in which the mother, neighbor and daughter nodes are all connected to each other [10]. The directed ND model, however, captures some essential properties of scientific citation networks, modeling the fact that a citation of a paper is often accompanied by citations to some of the earlier papers that appear in its reference list [11].

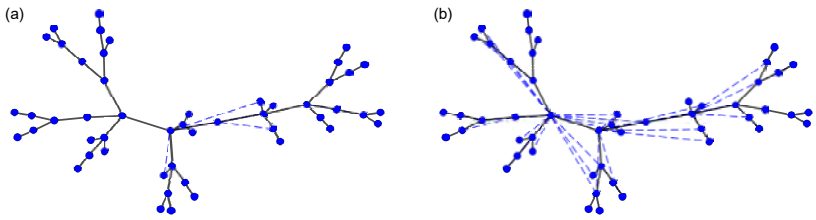


Fig. 1. Two instances of undirected corded ND networks of size $N = 50$, with $p = 0.1$ (a) and $p = 0.4$ (b). Both instances are formed around the same backbone tree (solid lines). Increasing p makes the network denser.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

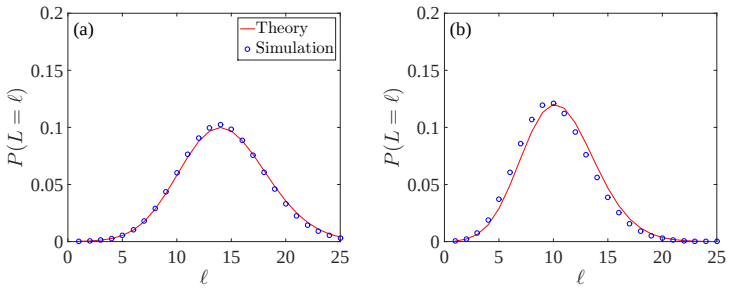


Fig. 2. The DSPL of the corded ND network of $N_t = 10^4$ nodes with (a) $p = 0.1$ and (b) $p = 0.3$. The theoretical results (solid lines) agree with the results of computer simulations (circles).

Recently, we obtained exact analytical results for the distribution of shortest path lengths (DSPL) in the corded ND undirected [12] and directed [8] models. To this end we derive master equations for the time evolution of the probability $P_t(L = \ell)$, $\ell = 1,2,\ldots$, where L is the distance between a random pair of nodes and t is the time. Note that the size of the network at time t is given by $N_t = s+t$, where s is the size of the seed network. Finding exact analytical solutions of the master equations, we obtain closed form expressions for $P_t(L = \ell)$. An important difference between the two cases is that while all pairs of nodes in the network are connected (or reachable from each other) in the undirected case, only a small fraction of pairs are reachable via directed links in the directed case. Surprisingly, the mean distance in both cases is found to scale logarithmically with the network size: namely $\langle L\rangle_t \sim \ln N_t$ in the undirected case, while the mean conditioned on reachable pairs $E_t[L|L < \infty] \sim \ln N_t$, thus the ND networks are small world networks. This result is in contrast to a common belief in the community stating that in scale-free networks the mean distance scales as $\langle L\rangle_t \sim \ln\ln N_t$ [13] - a behaviour known as "ultra small world". Actually, typical distances exhibited by our ND networks are much longer than those obtained in configuration model networks with precisely the same degree distribution. In Fig. 2 we present the distribution $P_t(L = \ell)$ for an ensemble of the undirected corded ND networks of size $N_t = 10^4$, grown from a seed network of size $s = 2$, with $p = 0.1$ and 0.3. The analytical results are found to be in excellent agreement with the results of computer simulations. The distribution turns out to be much broader than in other random networks (the directed case exhibits even a better agreement). In Fig. 3 we present the mean distance as a function of the network size N_t for both models and representative values of p. The results confirm the logarithmic scaling with network size.

In summary, we obtained exact analytical results for the distribution of shortest path lengths in corded node duplication networks. These results provide insight on the large scale structure of node duplication networks and on the relation between the growth process and the resulting structure. In particular, one important conclusion from this work is that the corded ND networks are small world networks rather than ultrasmall. This means that the large scale structure of networks is far richer than any prediction based on the degree distribution alone. Furthermore, The corded directed ND network

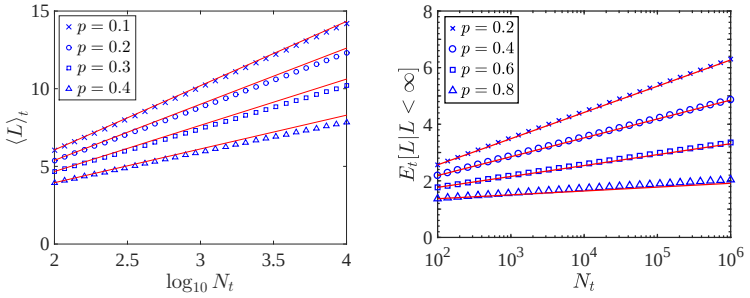


Fig. 3. The mean shortest path length, $\langle L \rangle_t$ (undirected) and $E_t[L|L < \infty]$ (directed), of ND networks as a function of network size N_t. The theoretical results (solid lines) confirm the logarithmic dependence on the network size. As p is increased, distances decrease in both cases.

provides some insight on the structure of the scientific citation networks. It indicates that for a given paper, the typical number of papers that are reachable to it by directed paths of citations (in the past or future) scales like $\ln N/N$. This provides an insight into why the scientific literature is highly fragmented in the sense that most pairs of papers are not reachable via chains of citations. For those pairs of papers that are reachable by chains of subsequent citations, the DSPL provides the breakdown into direct citations, indirect citations via a single intermediate paper and indirect citations of higher orders. This sheds new light on the way the impact of a paper may be evaluated, namely not only in terms of the direct citations but also in terms of the cumulative effect of all the secondary citations. Another aspect revealed by the model is that the structure of citation networks evolves slowly, with a typical logarithmic time-scale. However, it should be emphasized that this model is only a minimal model of citation networks. In more complete models a new paper may cite several 'mother nodes' as well as some of the earlier papers cited in them. This would increase the number of directed paths but is not expected to change the qualitative properties of the network.

References

1. A. Bhan, D.J. Galas and T.G. Dewey, *Bioinformatics* **18**, 1486 (2002).
2. J. Kim, P.L. Krapivsky, B. Kahng and S. Redner, *Phys. Rev. E* **66**, 055101 (2002).
3. F. Chung, L. Lu, T.G. Dewey and D.J. Galas, *J. Comput. Biol.* **10**, 677 (2003).
4. G. Bebek et al., *Theor. Comput. Sci.* **369**, 239 (2006).
5. S. Li, K.P. Choi and T. Wu, *Theor. Comput. Sci.* **476**, 94 (2013).
6. R. Lambiotte, P. L. Krapivsky, U. Bhat and S. Redner *Phys. Rev. Lett.* **117**, 218301 (2016).
7. U. Bhat, P. L. Krapivsky, R. Lambiotte and S. Redner *Phys. Rev. E.* **94**, 062302 (2016).
8. C. Steinbock, O. Biham and E. Katzav, *J. Stat. Mech.* 083403 (2019).
9. R. Toivonen et al., *Social Networks* **31**, 240 (2009).
10. M. Granovetter, *American Journal of Sociology* **78**, 1360 (1973).
11. M. Golosovsky, S. Solomon, *Phys. Rev. Lett.* **109**, 098701 (2012)
12. C. Steinbock, O. Biham and E. Katzav, *Phys. Rev. E***96** , 032301 (2017).
13. R. Cohen and S. Havlin, *Phys. Rev. Lett.* **90**, 058701 (2003).
14. C. Steinbock, O. Biham and E. Katzav, *Eur. Phys. J. B* **92**, 130 (2019).

Are degree distributions in complex networks observable?

Igor E. Smolyarenko

Brunel University London, Uxbridge, UB8 3PH, UK,
igor.smolyarenko@brunel.ac.uk

Introduction. Given an empirically observed complex network, a typical first step in analyzing its structure is to harvest node degrees and attempt to determine the law governing their distribution (assuming the network is large enough to perform a meaningful statistical analysis). A power-law degree distribution is frequently associated with a 'small-world' structure, thus allowing a quick glimpse into the network topology.

The prevalent paradigm in network science is that power-law degree distributions are ubiquitous [1, 2]. This, in turn, governs a lot of theoretical effort aimed at modelling complex networks. The opposite view has also been expressed [3, 4], in particular Ref. [4] generating a robust exchange of views [5–7]. Underlying this discussion is a view that application of suitably elaborate statistical methods is sufficient to answer this question one way or another. The crucial, and often unstated, assumption, is that one could employ standard *universal* statistical tools like Kolmogorov-Smirnov (KS) statistic [8], or other methods (*e.g.* Anderson-Darling, Cramér-von Mises) based on various measures of distance between the hypothesized and the empirical cumulative distribution functions (CDFs), in order to determine the 'goodness' of a degree distribution fit [4, 9]. The purpose of this note is to draw attention to the fact that an overlooked subtlety of the statistics of empirical CDFs of node degrees in networks may render the question of determining degree distribution laws conceptually undecidable.

Let us re-state briefly the textbook foundations of the KS and related methods. Consider a sequence of i.i.d. random variables $\{X_i\}_{i=1}^N$ characterised by CDF $F(x)$. We have $F_{emp}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{X_i \le x\}}$, hence $\mathbb{E}[F_{emp}(x)] = F(x)$, and $N^2\mathbb{E}[F_{emp}(x)F_{emp}(x')] = \sum_{ij}^{N}\mathbb{E}[\mathbf{1}_{\{X_i\le x\}}\mathbf{1}_{\{X_j\le x\}}] = \sum_{i=1}^{N}\min(F(x),F(x')) + \sum_{i\neq j}F(x)F(x')$. We thus obtain $\frac{1}{N}\min(F(x),F(x')) + \frac{N-1}{N}F(x)F(x')$, and therefore $\mathrm{Cov}[F_{emp}(x)F_{emp}(x')] = \frac{1}{N}\min(F(x),F(x')) + \frac{N-1}{N}F(x)F(x') - F(x)F(x') = \frac{1}{N}\{\min(F(x),F(x')) - F(x)F(x')\}$. Denoting $\mathscr{F}_{emp}(u) = F_{emp}(F^{-1}(u))$, one finds the universal scaling form $\mathscr{C}(u,v) \equiv N\mathrm{Cov}[\mathscr{F}_{emp}(u),\mathscr{F}_{emp}(v)] = \min(u,v) - uv$, independent of the specific form of $F(x)$. Universal applicability of the KS statistic follows from the convergence of the normalised empirical CDF $\mathscr{F}_{emp}(u)$ to Brownian bridge on $u \in [0,1]$, characterised by the covariance function above.

A crucial element in this calculation is the cancellation of $O(1)$ terms generated by factorization of $\mathbb{E}[\mathbf{1}_{X_i\le x}\mathbf{1}_{X_j\le x}]$. It follows that in the case of *dependent* variables very weak correlations leading to an $O(1/N)$ correction to factorization of this term are enough to produce an $O(1)$ contribution to the universal scaled covariance function, rendering the statistics of KS distances non-universal.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Degree distribution of a random network is a *marginal* distribution derived from the 'global' distribution of the full adjacency matrix, hence *a priori* there is no reason why such correlations would not be present, breaking the applicability of *any* testing method relying on a measure of distance between empirical and conjectural CDFs to degree distributions. Note that at issue is *not* the usual measure of degree correlations in networks focusing on pairs of adjacent nodes [10], but a more 'egalitarian' measure of correlations among arbitrarily chosen pairs of nodes. We consider three paradigmatic (undirected) network models: (i) Erdős-Rényi (ER) networks, (ii) Barabási-Albert (BA) networks, and (iii) inhomogeneous ER networks (iER) and static fitness networks [11, 12], geared towards producing power-law degree distributions. In all cases we show that covariance function of F_{emp} is finitely different from the *i.i.d.* case.

ER networks. We define an $N \times N$ ER network as a set of *i.i.d.* Bernoulli random variables S_{ij}, where $S_{ij} = 1$ if nodes i and j are connected (with probability p), and $S_{ij} = 0$ otherwise. We therefore have $F_{emp}(k) = \frac{1}{N}\sum_{\kappa=0}^{k} n_\kappa$, where $n_\kappa = \sum_{i=1}^{N} \mathbf{1}_{\{d_i=\kappa\}}$ is the total number of nodes with degree $d = \kappa$. In the dense network regime $p \sim O(1)$, we find the standard result $F(k) \equiv \mathbb{E}[F_{emp}] = \Phi\left([k-\bar{k}]/\sqrt{N-1}\sigma\right)$, where $\bar{k} = pN$, $\sigma^2 = p(1-p)$, and $\Phi(z)$ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Extending the analysis to degree correlations, we obtain

$$\mathrm{Cov}[\sqrt{N}\mathscr{F}_{emp}(u), \sqrt{N}\mathscr{F}_{emp}(v)] = \mathscr{C}(u,v) + \Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(w)) \qquad (1)$$

The exact functional form of the correction term is borne out by numerical simulations (Figure 1). Similar results are obtained in the case $p \sim O(1/N)$.

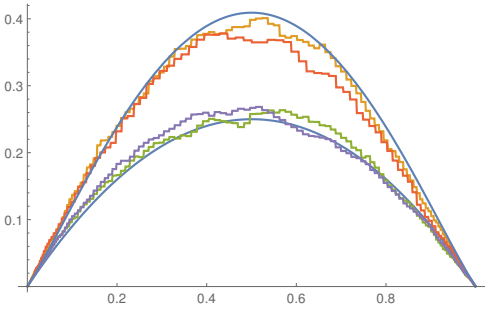


Fig. 1: Scaled variance of F_{emp} for different values of p using either independent Gaussian variables (lower cluster of curves), or degree distribution of ER graphs, overlaid with the plot of $u - u^2 + \Phi'^2(\Phi^{-1}(u))$ (upper cluster of curves), in accordance with the diagonal limit $u = v$ of Eq. (1).

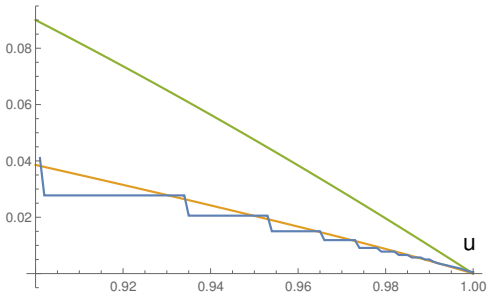


Fig. 2: The simulated variance of F_{emp} for BA graphs, overlaid by the theoretical curve $u - u^2 + S(k(u))$, and the Brownian bridge variance $u - u^2$ (top curve) for comparison. Only the rightmost part of the plots is shown, as the bulk of the nodes in BA networks have low degrees.

BA networks. A standard approach using rate equations [10] can be extended to the analysis of $P(k,k';N) = \frac{1}{N(N+1)} \sum_{\substack{\tau=0,\tau'=0 \\ \tau\neq\tau'}}^{N} \mathbb{P}[\mathrm{d}(N|\tau) = k \cap \mathrm{d}(N|\tau') = k']$, where $\mathrm{d}(N|\tau)$ is the degree (at discrete time N) of the node preferentially joined to the network at some previous time τ. The full covariance function of F_{emp} is rather cumbersome, so we present here the result at coinciding arguments: $N\mathrm{Var}[\mathscr{F}_{emp}(u)] = u - u^2 + S(k(u))$, where the correction term is $S(k) = -\frac{2k^2(7+2k)}{(k+1)(k+2)(2k+1)} + 2F^2(k)$, the CDF of the infinite BA network is $F(k) = \frac{3k+k^2}{(k+1)(k+2)}$, and $k(u)$ is the corresponding inverse function. The exact functional form of the correction term is again borne out by numerical simulations (Figure 2). Note that in contrast to the ER case, the variance of $\mathscr{F}_{emp}$ here is *smaller* than for independent random variables.

iER and static fitness networks. We now consider iER networks with $\mathbb{P}[S_{ij} = 1] = (\lambda/N < w >_w) w_i w_j$, where the sequence w_i is chosen [13] so that the degree distribution is asymptotically a power law $p(k) \sim k^{-\beta}$, λ controls the overall density of links, and $< \phi(w) >_w = \sum_i \phi(w_i)/N$. Denoting $f(k,w) = w^k e^{-w}/k!$, and $F(k,w) = \sum_{\kappa=0}^{k} f(\kappa,w) = \Gamma(k+1,w)/\Gamma(k+1)$, where $\Gamma(n,z)$ is the (upper) incomplete Γ-function, we find

$$N\mathrm{Var}[\mathscr{F}_{emp}(u)] = \left\langle F(k(u),\lambda w) - F^2(k(u),\lambda w) \right\rangle_w + \left(\frac{\lambda}{< w >_w} \right) \langle w f(k(u),\lambda w) \rangle_w^2 \tag{2}$$

where $k(u)$ is the inverse function of $F(k) = \langle F(k,\lambda w) \rangle_w$.

Figure 3 (bottom curves) shows that in the iER model (choosing $\beta = 3$ to match the BA case), empirical variance is finitely smaller than the independent case, and displays a particularly strong suppression near $F = 1$. The top curve in Figure 3 shows numerically simulated variance for fitness-based [11, 12] tree networks, with $\beta = 3$. Note that the variance in this case is *larger* than in the independent case, despite the same $\beta = 3$ power-law degree distribution as in the BA and iER networks. [Theory for the variance of F_{emp} is not yet available in the fitness model.]

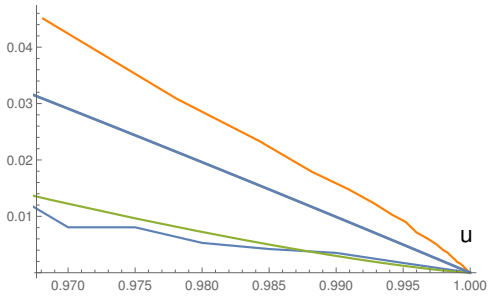


Fig. 3: The simulated variance of F_{emp} for iER graphs, overlaid by the theoretical curve obtained from Eq. (2) for $\beta = 3$, the Brownian bridge variance $u - u^2$ (middle curve) for comparison, and fitness model (top curve).

Conclusion. We demonstrate that variance of empirical CDF of network degree distributions is highly non-universal, depending on the details of each of the generative models explored. Crucially, we have shown that models with the same $\beta = 3$ power-law degree distribution may exhibit both suppression and enhancement of the variance

of F_{emp} compared to the case of independently distributed variables. Consequently, universal KS (or similar) test statistic cannot be applied. Furthermore, each empirically observed network is *sui generis*, with unknown, and most often unknowable, growth or creation mechanism. *It therefore appears impossible to use any bespoke bootstrapping methods to simulate the distribution of KS (or similar) test statistic for network degree distribution.* The fact that the effect is finite as $F \to 1$ means that methods based on tail estimators [6] may be similarly impacted. The results of this study naturally lead to pose the question whether the full information contained in a *single instance* of an observed adjacency matrix (thus abandoning the concept of degree distribution as a self-contained network characteristic) could be exploited to determine the sign of the deviation of the variance from the universal case, and hence open the way to some approximate bootstrapping approaches.

Acknowledgements. Discussions with E. Wit, V.Vnciotti, and I. Artico are gratefully acknowldged.

References

1. M. E. J. Newman, Contemporary Physics 46, 323 (2005), 0412004.
2. A. Barabasi, Network Science, Cambridge University Press, 2016.
3. R. Khanin and E. Wit. Journal of computational biology, 13, 810 (2006).
4. A. D. Broido and A. Clauset, Nature Communications 10, 1017 (2019).
5. A. Barabasi, Love is All You Need: Clauset's fruitless search for scale-free networks, available at https://www.barabasilab.com/post/love-is-all-you-need, 2018.
6. I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov. arXiv preprint arXiv:1811.02071, 2018.
7. P. Holme, Nature Communicationsvolume 10, 1016 (2019).
8. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2012.
9. A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review 51, 661 (2009).
10. P. L. Krapivsky and S. Redner. Organization of growing random networks. Physical Review E, 63, 066123 (2001).
11. G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. 89, 258702 (2002).
12. I. E. Smolyarenko, K. Hoppe, and G. J. Rodgers, Physical Review E 88, 012805 (2013).
13. F. Chang, L. Lu, V. Vu, Internet Math. 1, 257 (2003).

Optimal change point estimator for network data

Sharmodeep Bhattacharyya[1], Shirshendu Chatterjee[2], and Soumendu Sundar Mukherjee[3]

[1] Oregon State University, Corvallis OR 97331, USA,
Sharmodeep.Bhattacharyya@oregonstate.edu,
WWW home page: https://stat.oregonstate.edu/people/bhattacharyya-sharmodeep
[2] City University of New York, New York NY 10031, USA
shirshendu@ccny.cuny.edu,
WWW home page: shirshendu.ccny.cuny.edu
[3] Indian Statistical Institute, Kolkata 700108, WB India,
soumendu041@gmail.com,
WWW home page: https://soumendu041.gitlab.io/

1 Introduction

Change-point detection is a classical problem in statistics which has gained significant importance and applicability in many fields including medical diagnostics, gene expression, spam email filtering, astronomy and finance. Such problems arise in the analysis of various data types including sequentially observed normally distributed data, time-series data and multivariate data for detecting change in different parameters of the data distribution such as mean, variance, correlation, density.

In this article, we tackle the problem of change-point detection in temporal network data. The observable is a sequence of networks indexed by time. The goal is to check if there is any time-point, which will be referred to as change-point, when there is a significant change in the structure of these networks and to estimate the location of such change-points. These problems arise in many applications including (i) brain image analysis, where one observes scanned images of brains collected over time and looks for abnormalities, (ii) ecological networks observed over time, where one checks whether there is any structural change. We stress here that we observe the whole time series ahead of our analysis, this is thus an *offline* or *a posteriori* change-point problem.

Recent work in this area include [6] (Bayesian procedure for hierarchical random graph model), [5, **?**] (use of local graph statistics for anomaly detection in dynamic networks), [2] (eigenvalue based test to segregate graph models). Although much empirical work has been done, not much theory can be found (exception is [7]), and most theoretical results focus on particular structures or specialized models. Some recent works [8, **?**] propose methods for change point detection in networks generated from block models and graphon models with some theoretical results on the consistency of the detection methods.

The classical CUSUM statistic [4] for univariate change point problems can be used in the network problem as well, and provides a unified way of constructing estimates of change points. A preliminary study of its theoretical poperties was carried out in [3]. In this paper, we will further that investigation in a much more general setting under very mild assumptions.

2 Contribution of the paper

We consider the setup where one observe T networks with adjacency matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{T1)}$ on the same set of nodes $\{v_1, v_2, \ldots, v_n\}$, where the edges in $\mathbf{A}^{t1)}$ appear independently and $\mathbb{E}\mathbf{A}^{(t)} =$

P_t. The first problem that we address is testing the hypotheses

$$H_0 : P_i = P, 1 \le i \le T, \text{ versus } H_1 : \exists\, 1 \le \tau \le T-1 \text{ such that } P_i = \begin{cases} P & 1 \le i \le \tau \\ Q & \tau+1 \le i \le T. \end{cases}$$

and estimating τ when H_1 is true. Let $\tau \in (\kappa, T-\kappa)$ and Λ is the target precision for estimating τ. Also let $\bar{D}$ be the sample average degree of a node over all layers. Define

$$\text{For } \ell \in [T/\Lambda], \text{ let } T_\ell := (\ell-1)\Lambda \text{ and } \mathbb{J}_\ell := \cup_{\star\in\{+,-\}} \mathbb{I}_\star \left(T_\ell, T_{\ell+1}; \frac{1}{3}\Lambda \wedge \kappa \right),$$

$$\text{where } \mathbb{I}_-(a,b;c) := \{(a,t] : a+c \le t \le b\}, \mathbb{I}_+(a,b;c) := \{(t,b] : a \le t \le b-c\}$$

$$\Gamma := n \cdot \min \left\{ \frac{1}{2}, \left(\frac{\log(T/\Lambda)}{\bar{D}^3(\Lambda \wedge \kappa)} \right)^{1/2} \right\},$$

$$D_{i,\ell} := \max_{J \in \mathbb{J}_\ell} \frac{1}{|J|} \sum_{j\in[n], s\in J} A_{ij}^{(s)} \text{ for } i \in [n] \text{ and } \ell \in [T/\Lambda],$$

Algorithm 1: Change Point Detection
Input: Adjacency matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(T)}$; cushion κ, scanning window Λ.
Output: Change point estimate $\hat{\tau}$.

Obtain $\bar{D} = \frac{1}{nT} \sum_{i,j\in[n], s\in[T]} A_{ij}^{(s)}$.
Obtain Γ.
For $\ell = 1, 2, \ldots, T/\Lambda$ do
  Obtain T_ℓ and $\mathbb{J}_\ell$.
  For $i = 1, 2, \ldots, n$ do
    Obtain $D_{i,\ell}$.
  Order the values $D_{1,\ell}, \ldots, D_{n,\ell}$ to get $D_{(1),\ell} \le \cdots \le D_{(n),\ell}$.
  Obtain row indices $i_1, \ldots, i_\Gamma$ such that $D_{i_k,\ell} \ge D_{(n+1-\Gamma),\ell}$
  Obtain $\tilde{\mathbf{A}}^{(s)}$ from $\mathbf{A}^{(s)}$ for each $s \in (T_\ell, T_{\ell+1}]$ by removing rows and columns with indices $i_1, \ldots, i_\Gamma$.
  For $t = \frac{1}{3}(\Lambda \wedge \kappa), \frac{1}{3}(\Lambda \wedge \kappa) + 1, \ldots, \Lambda - \frac{1}{3}(\Lambda \wedge \kappa)$ do
    Obtain $\mathbf{G}^{(t)} := \frac{1}{t} \sum_{s\in(0,t]} \tilde{\mathbf{A}}^{(s+T_\ell)} - \frac{1}{\Lambda - t} \sum_{s\in(t,\Lambda]} \tilde{\mathbf{A}}^{(s+T_\ell)}$.
  Obtain $u = \arg\max_{t\in(T_\ell, T_{\ell+1}]} \left\| \mathbf{G}^{(t)} \right\|$
  If $\left\| \mathbf{G}^{(u)} \right\| > C\Psi \left[\frac{\bar{D}}{(\log n)\Lambda \wedge \kappa} \log(CT/\Lambda) \right]^{1/2}$, declare u as a change point.

A natural statistic to consider under our single change-point alternative is based on the cumulative averages of estimates of the P_is. Such CUSUM statistics are very widely used in change-point problems. We use $\mathbf{A}^{(t)}$ as an estimate of P_t. Then we obtain submatrices $\tilde{\mathbf{A}}^{(t)}$ of $\mathbf{A}^{(t)}$ by removing some high degree vertices as described in Algorithm 1. Then we obtain

$$G_t := \frac{1}{t} \sum_{i=1}^{t} \tilde{\mathbf{A}}^{(i)} - \frac{1}{T-t} \sum_{i=t+1}^{T} \tilde{\mathbf{A}}^{(i)} \text{ for } \kappa \le t \le T-\kappa.$$

We accept H_0 (no change-point) if $\max_{\kappa \le t \le T-\kappa} ||G_t|| \le C\sqrt{\bar{D}/T}$, where $\bar{D}$ is the estimated average degree, $||\cdot||$ denotes the spectral norm and C is a large constant. Otherwise, we accept H_1 (existence of a change-point) and obtain $\hat{\tau}$ (our estimate of the single change-point τ) by

$$\hat{\tau} := \arg \max_{\kappa \le t \le T-\kappa} ||G_t||.$$

In many univariate settings, such CUSUM statistics are minimax optimal (see, e.g., [1]).

Theorem 1. *Let d be the average degree of a node among the networks $A_1, \ldots, A_T$. Then $|\hat{\tau} - \tau| = o(T)$ when $||P - Q|| \gg \sqrt{d/T}$. Also, if $||P - Q|| \not\gg \sqrt{d/T}$, detection is not be possible.*

We also address the case of multiple change-points, where there are K (unknown) change-points $\tau_1 < \tau_2 < \cdots < \tau_K$ and obtain consistent estimates $\hat{K}, \hat{\tau}_1, \ldots, \tau_K$ under minimal assumptions on network parameters.

Comment: Note that, Theorem 2.1 states the optimal condition on operator norm of the difference between the connection probability matrices, $||P - Q||$ for recovery of change-point. The optimal condition depends on $\sqrt{d/T}$.

Summary. Based on a finite sequence of observed independent network data, we provide an algorithm to test if there is any change-point and estimate the location of the change-points (if any). The algorithm works effectively whenever the signal (norm of the difference of means of the networks before and after the change-point) is above the detectability threshold irrespective of whether the networks are sparse or dense.

References

1. E. Brodsky and B. S. Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media, 2013.
2. J. Cape, M. Tang, and C. E. Priebe. The kato–temple inequality and eigenvalue concentration with applications to graph inference. *Electronic Journal of Statistics*, 11(2):3954–3978, 2017.
3. Soumendu Sundar Mukherjee. *On Some Inference Problems for Networks*. PhD thesis, May 2018.
4. E. S. Page. On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, 44(1/2):248–252, 1957.
5. Y. Park, C. E. Priebe, and A. Youssef. Anomaly detection in time series of graphs using fusion of graph invariants. *IEEE journal of selected topics in signal processing*, 7(1):67–75, 2013.
6. L. Peel and A. Clauset. Detecting change points in the large-scale structure of evolving networks. In *AAAI*, pages 2914–2920, 2015.
7. S. Roy, Y. Atchadé, and G. Michailidis. Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1187–1206, 2017.
8. Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*, 2018.

Part XIII

Network Neuroscience

Slow and Anomalous Dynamics in Hierarchical Modular models of Brain Networks

Samaneh Esfandiary[1], Ali Safari[1], Miguel A. Muñoz[2], and Paolo Moretti[1]

[1] Institute for Materials Simulation, Fredrich-Alexander-University Erlangen-Nuremberg
Dr-Mack-Str 77 90762 Fürth, Germany,
samaneh.esfandiary@fau.de,
WWW home page: http://www.matsim.techfak.uni-erlangen.de
[2] Departamento de Electromagnetismo y Fsica de la Materia e Instituto Carlos I de Fsica Terica y Computacional, Universidad de Granada, ,
Granada E-18071, Spain

1 Introduction

Activity in the human brain displays features of localization, allowing for multiple tasks at the same time, and slow relaxation, which means that in localized regions activity is sustained for long times. This phenomenology can be explained starting from the structure of brain connectivity patterns, which is that of a hierarchical modular network (HMN) [1–3]. In this work we address the study of these two features, localization and anomalous dynamics. To achieve this objective, we use a minimal computational model for diffusion, as a simple dynamic protocol, able to capture the essence of the dynamic slowing down due to the hierarchical organization. Diffusion and random walk simulations are tightly related to the concept of spectral dimension of a network. Hierarchical modular networks are known to have a finite topological dimension. This property has been variously invoked in the past to explain the emergence of localization and the dynamic slowing down in such networks [3, 4]. The topological dimension remains however a purely structural measure, and its role in generating localized functional patterns such as rare-region events has been only conjectured. The spectral dimension, instead, can be computed measuring the return probability of a simple random walk simulation, providing a direct connection between structure and function (diffusive dynamics in this case). Our approach thus consists in running vary-large-scale simulations (network sizes up to 30 million nodes), for the accurate computation of spectral dimension in HMNs. Our results show that the dynamic slowing down in such systems can be related to the surprising fact that the spectral dimension for such systems is undefined, i.e. the return probability for HMNs does not decrease in time as a simple power law, and it rather exhibits a stretched exponential correction. Since several brain pathologies are associated with a dimension reduction and/or anomaly (as in a breaking process), our results may serve as the foundation for topology-based diagnostic tools, in the broader context of network physiology and medicine [5].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

For simplicity, we construct HMNs using the method introduced in [3] and we call α the connectivity strength of the HMN, being α proportional to both the average network degree $\langle k \rangle$ and (asymptotically) to its topological dimension [6].

As we are interested in computing the spectral dimension D_s of such networks, we recall that D_s can be measured in random walk simulations. To this end, we define the average return probability $P_{ii}(t)$ as the probability of a walker to return to its starting node i after t steps (at time t), averaged over all choices of i. If P_{ii} behaves asymptotically as

$$P_{ii}(t) \sim t^{-\frac{D_s}{2}}, \tag{1}$$

then D_s is the spectral dimension of the network. In any other case, the spectral dimension is not defined, thus pointing to anomalous (and often very slow) relaxation. We compute D_s by running large-scale random walk simulations, extracting P_{ii}, and verifying if the standard behavior in 1 holds. Our findings are exemplified in Figure 1, where we show a sample curve of the return probability for a HMN. While the exponential cutoff at extremely large values of t is the expected cutoff associated with the network's finite size, the initial power-law-like behavior depends non-trivially on the choice of α and asymptotically exhibits a stretched exponential correction at very large t, $P_{ii} \sim \exp\left[-(t/t_0)^{\beta}\right]$ pointing to an excess of return events at very large times. The exponent $0.5 < \beta < 1$ is found to depend on the connectivity strength α: lower values of α (and lower topological dimension) produce lower values of β (and more significant slowing down). For higher α instead, β approaches 1 and the standard behavior is recovered. The details of our analysis of the anomalous exponent β are shown in Figure 2, which allows us to conclude that the standard diffusive scenario associated with networks of finite spectral dimension does not hold in the case of HMNs. We note that the initial power-law-like regime in Figure 1 depends surprisingly weakly on α and the analysis of its role in HMNs will be the subject of future work.

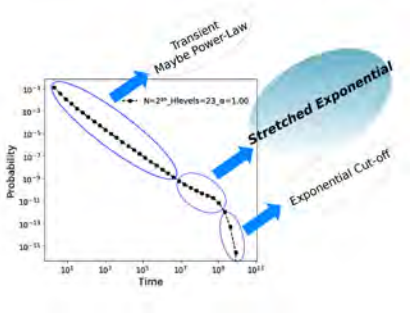


Fig. 1. Sample curve of the return probability of a random walk on HMNs. The initial power-law-like regime crosses over to a slower regime, whose scaling is analyzed in Figure 2. The final exponential cutoff is due to the finite size of the network.

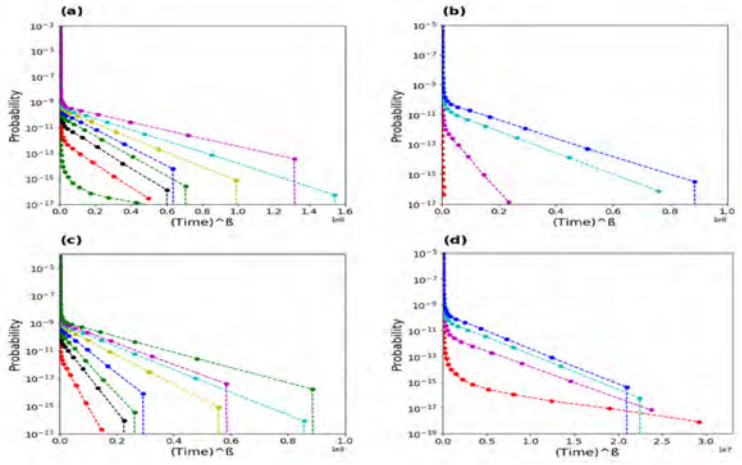


Fig. 2. Analysis of the stretched exponential regime, for HMNs of varying geometries, and for increasing values of α from bottom to top. (a) $N = 2^{25}$, 23 hierarchical levels; (b) $N = 2^{25}$, 24 hierarchical levels; (c) $N = 2^{24}$, 22 hierarchical levels; (d) $N = 2^{24}$, 23 hierarchical levels.

Summary. We show how the standard diffusive scenario for finite dimensional networks does not hold for hierarchical modular network models of the brain, where anomalous slowing down phenomena are encountered. We arrive at our conclusion by computing the spectral dimension in networks of sizes up to 30 million nodes. We observe that the asymptotic behavior of the return probability points to anomalous dynamic patterns occurring at very large times, making the standard measure of spectral dimension undefined. This anomalous behavior is due to structural features of the network and may contribute to the ability of brain networks to host localized patterns of activity and conduct multiple tasks at the same time.

References

1. Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C J., Wedeen, V J., Sporns, O.: Mapping the structural core of human cerebral cortex. PLoS Biol. 6(7), e159 (2008)
2. Sporns, O., Chialvo, D. R., Kaiser, M., Hilgetag, C. C. Organization, development and function of complex brain networks. Trends in cognitive sciences 8(9), 418-425(2004).
3. Moretti. P and Muñoz M. A.: Griffiths phases and the stretching of criticality in brain networks. Nature Commun. 2521, 4 (2013)
4. Gallos, L. K., Makse, H. A., Sigman, M.: A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. Proc. Natl. Acad. Sci. USA 109, 2825-2830 (2012)
5. Ivanov, P. C., Liu, K. K. L., Bartsch, R. P.: Focus on the emerging new fields of network physiology and network medicine, New J. Phys. 18, 100201 (2016)
6. Safari, A., Moretti, P., Muñoz, M. A. Topological dimension tunes activity patterns in hierarchical modular networks. New J. Phys. 19, 113011 (2017)

Recurrence Analysis of Dynamic Brain Networks: Characterisation of the Spatio-Temporal Dynamics of Magnetoencephalographic Recordings

Marinho A. Lopes[1,2], Jiaxiang Zhang[2], Dominik Krzemiński[2], Khalid Hamandi[2], Lorenzo Livi[3,4], and Naoki Masuda[1,5]

[1] Department of Engineering Mathematics, University of Bristol BS8 1UB, United Kingdom
m.lopes@bristol.ac.uk
[2] Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, Cardiff CF24 4HQ, United Kingdom
[3] Departments of Computer Science and Mathematics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
[4] Department of Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, United Kingdom
[5] Department of Mathematics, University at Buffalo, State University of New York, USA
[6] Computational and Data-Enabled Science and Engineering Program, University at Buffalo, State University of New York, USA

1 Introduction

The brain is a complex network whose function results from multiscale spatio-temporal dynamics. Techniques such as electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) allow us to record brain activity and give us a window into brain function [1]. Traditional approaches to studying brain function (and dysfunction) have primarily focused on the structure of static brain networks [2]. However, functional brain networks are dynamic because they depend on time-evolving brain activity (see e.g. [3]).

Recurrence plots (RP) have been introduced in the 1980's to study time series from dynamical systems. From the exploration of their properties a number of measures were developed and recurrence quantification analysis (RQA) emerged [4]. This approach has been widely used to investigate brain activity. For example, Ngamga *et al*. [5] used RQA to identify pre-seizure states from intracranial EEG data recorded from people with epilepsy. Such approaches have neglected however the spatial correlations between the brain activity recorded from different brain areas.

We propose to employ RQA on dynamic functional brain networks to study spatio-temporal brain dynamics. Herein we apply this approach to a resting-state MEG dataset comprising 26 people with epilepsy and 26 controls. Our purpose is to illustrate the application of our framework and to explore whether such methodology may unveil biomarkers of epilepsy.

2 Methods

Resting-state MEG data was recorded from 26 people with juvenile myoclonic epilepsy and 26 healthy controls. The MEG data was filtered in the classical frequency bands

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

(theta (4-7 Hz), alpha (8-13 Hz), beta (15-25 Hz) and gamma (30-60 Hz) bands), and source-reconstructed using a linear constrained minimum variance (LCMV) beamformer on a 6-mm template with a local spheres forward model. Sources were mapped into the 90 brain regions of the Automated Anatomical Label (AAL) atlas [6].

To build time-dependent functional networks, we divided the MEG recordings into segments of 500 sample points (2 seconds) with 80% overlap between consecutive segments. We considered 500 segments per individual. For each segment, we constructed functional networks using two well-established methods: phase lag index (PLI) [7] and amplitude envelope correlation (AEC) [6].

To apply RQA to dynamic functional networks it is first necessary to build RPs. A RP is represented by a binary recurrence matrix $\mathbf{R} = (R_{i,j})$, where $R_{i,j} = 1$ if the functional network at time point i is within a small distance to the functional network at time j; otherwise $R_{i,j} = 0$. We explored a number of different possible distance measures between networks: Frobenius norm, spectral norm, and log-Euclidean norm of the difference between adjacency matrices. Furthermore, we also considered distance measures based on the networks' Fiedler vectors: cosine dissimilarity, Euclidean norm, and infinity norm between the Fiedler vectors. To define the thresholds of recurrence, i.e., the distance within which $R_{i,j} = 1$, we imposed a fixed density of recurrence points in the RP. In particular, we considered thresholds such that the density was equal to 0.01, 0.05, and 0.10. We then used the Cross Recurrence Plot Toolbox to compute the RQA [4, 8]. Figure 1 illustrates the key steps of our method.

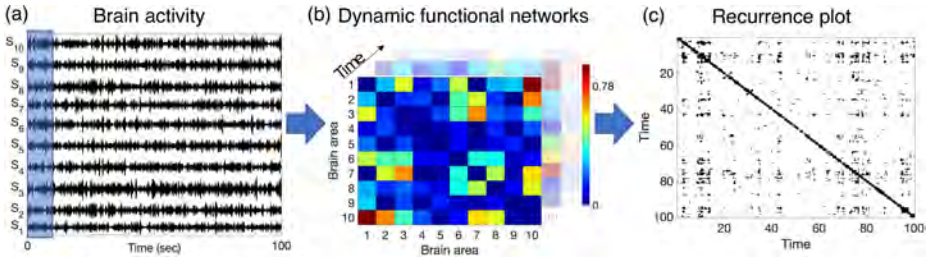


Fig. 1. Scheme of the data analysis procedure to apply RQA to dynamic functional networks. (a) Brain activity is segmented into windows. (b) From each window, a functional brain network is inferred. (c) A recurrence plot (RP) is computed by assessing the distance between functional networks at different times. Black dots correspond to pairs of functional networks that were within a distance smaller than a threshold. RQA allows us to then extract information from the RP.

In this study, we consider four frequency bands, two functional network measures, and six distance measures. We define a *configuration* as one combination of frequency band, functional network and distance measure. For example, theta-AEC-Frobenius norm is one configuration. Thus, we have $4 \times 2 \times 6 = 48$ different configurations. However, different configurations may yield similar information when we apply RQA. Therefore, we first studied the relations between the 48 configurations using principal component analysis, k-means clustering, and Pearson correlation. This investigation aimed at removing redundant configurations from further analysis. We then performed

RQA with a limited number of configurations: we used Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons to assess different recurrence measures between people with epilepsy and controls.

3 Results

From a preliminary analysis, we found four sufficiently independent configurations of methodological choices to study the MEG data (AEC networks combined with Frobenius norm, spectral norm, and Euclidean distance between the Fiedler vectors, as well as PLI networks combined with spectral norm). All frequency bands offered similar information, and thus we focused our analysis on the theta band.

We observed statistically significantly smaller recurrence times (of both 1st and 2nd types [4]) in people with epilepsy compared to controls. This implies that functional networks from people with epilepsy are more likely to recur more often than those from controls.

4 Conclusions

RQA applied to dynamic functional networks can be used to reveal spatio-temporal features of brain activity. In particular, when applied to resting-state MEG data from people with epilepsy, it reveals that their functional networks recur more often in time than those from healthy controls. This suggests that epilepsy may reduce the brain's repertoire of functional states. In future work we aim to examine whether the recurrence time of dynamic functional networks may be used as a biomarker of cognitive impairment.

References

1. Buzsaki, G.: Rhythms of the Brain. Oxford University Press (2006)
2. Fornito, A., Zalesky, A., Bullmore, E.: Fundamentals of brain network analysis. Academic Press (2016)
3. Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., Grafton, S. T.: Dynamic reconfiguration of human brain networks during learning. Proc. Natl. Acad. Sci. U.S.A. 108(18), 7641–7646 (2011)
4. Marwan, N., Romano, M. C., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. Phys. Rep. 438(5-6), 237-329 (2007)
5. Ngamga, E. J., Bialonski, S., Marwan, N., Kurths, J., Geier, C., Lehnertz, K. Evaluation of selected recurrence measures in discriminating pre-ictal and inter-ictal periods from epileptic EEG data. Phys. Lett. A 380(16), 1419-1425 (2016)
6. Hipp, J. F., Hawellek, D. J., Corbetta, M., Siegel, M., Engel, A. K.: Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat. Neurosci. 15(6), 884 (2012)
7. Stam, C. J., Nolte, G., Daffertshofer, A.: Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. Hum. Brain Mapp. 28(11), 1178-1193 (2007)
8. Cross Recurrence Plot Toolbox: http://tocsy.pik-potsdam.de/CRPtoolbox/

The role of modularity in the formation of macroscopic patterns on functional brain networks

Bram A. Siebert[1], Malbor Asllani[1], Cameron L. Hall[1,2], and James P. Gleeson[1]

[1] MACSI, University of Limerick, Limerick V94 T9PX, Ireland
[2] Department of Engineering Mathematics, University of Bristol, Bristol, BS8 1UB, UK

1 Introduction

Emergent patterns of neuronal activity are key to a range of processes in the brain including brain development [1, 2], and neurological conditions [3]. Experimental studies have mainly focused on brain network topology, and it would be useful to understand the connection between this and the observed neuronal activation patterns. Using fMRI techniques, researchers have found that macroscopic patterns (patterns where entire modules of nodes synchronise) appear in brain networks [2, 4, 5]. Turing theory, originally developed to study spatial patterns [6], has been applied to pattern formation on networks since the 1970s [7, 8]. Turing instabilities (and the related pattern formation) can only occur when there is a sufficiently large difference between the rates of diffusion of two chemicals: fast-diffusing activators and slow-diffusing inhibitors. In this work, we study how modularity, a key feature of brain network topology, contributes to the ease with which Turing instabilities can occur in the Fitzhugh–Nagumo system, a widely studied model of neuronal activation. This builds on previous work on pattern formation on directed networks and multiplex networks [9, 10] We find that modularity enables pattern formation to occur in cases where the ratio between the activator diffusion coefficient and the inhibitor diffusion coefficient is much closer to one than would permit pattern formation on networks with other topologies, such as small-world networks.

2 Results

Our main result is that modular networks can exhibit Turing instabilities for cases where the ratio between the activator diffusion coefficient and the inhibitor diffusion coefficient is close to one. In this sense, the topology of brain networks is well-suited to the development of Turing patterns. From [6], Turing patterns can arise in nonlinear systems of reaction-diffusion equations involving at least two species. On a network, such systems take the form

$$\begin{aligned} \frac{\partial u_i}{\partial t} &= f(u_i, v_i) + D_u \sum_j \mathbf{L}_{ij} u_j \\ \frac{\partial v_i}{\partial t} &= g(u_i, v_i) + D_v \sum_j \mathbf{L}_{ij} v_j, \end{aligned} \tag{1}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where u_i is the concentration of activator in node i, v_i is the concentration of inhibitor in node i, $\mathbf{L}$ is the graph Laplacian, and D_u and D_v are diffusion constants. The functions f and g represent the net production of activator and inhibitor respectively; in our work, we use the well-established Fitzhugh–Nagumo equations for f and g throughout. Turing instabilities arise when a spatially-homogeneous steady state is linearly unstable to spatially-inhomogeneous perturbations. This can only occur when there is a sufficiently large difference between D_u and D_v. This causes the initially homogeneous steady state to evolve to a stable inhomogeneous state with different amounts of activator and inhibitor on each node: a Turing pattern, as in Fig. 1a.

Turing instabilities are studied by linearising about the homogeneous steady state and calculating the dispersion relation; in the network case, this relates the eigenvalues of the Laplacian (which represent different spatial patterns of instability via their associated eigenvectors) to the eigenvalues of the Jacobian (which represent the speed at which a small perturbation will grow or shrink). Instabilities will occur if and only if there is at least one eigenvalue of the Jacobian with a positive real part.

However, observations of natural systems show that D_u/D_v is often closer to 1. It can be shown that as $D_u/D_v \to 1$ the range of Laplacian eigenvalues corresponding to positive Jacobian eigenvalues shifts to be closer to 0. This makes it important to study Turing instabilities on systems with a small spectral gap, the distance between the zero eigenvalue and the first non-zero one of the Laplacian matrix. Brain networks manifest a modular organisation structure which induces a small spectral gap. They are structured in a hierarchical way where the whole network is organised in modules, which in turn are small-world graphs [11, 12]. This way, brain networks can still keep a good level of communicability due to the locally short distances, but at the same time can self-organise in spatially extended patterns due to the small spectral gap induced by the modularity. This is a key factor in why a modular organisation is important for the formation of patterns.

Turing patterns on modular networks also have the property of being macroscopic. This is because the solutions of the linearised system above are $\delta u_i = \sum_\alpha c_\alpha e^{\lambda_\alpha t} \Phi_i^\alpha$, where λ^α are the eigenvalues of the Laplacian, Φ^α are the associated eigenvectors, and c_α depends on the initial conditions. We can plot the normalised eigenvector corresponding to the largest eigenalues of the Jacobian together with the normalised steady-state pattern and show that the two can match well, see Fig. 1c. When D_u/D_v is close to 1, the largest eigenvalue of the Jacobian will correspond to the modular Laplacian eigenvalues, the first $m-1$ negative eigenvalues, where m is the number of modules. The eigenvectors corresponding to these modular eigenvalues have the property that nodes in the same module are associated with similar eigenvector values. This has the overall effect of leading to macroscopic Turing patterns.

Summary Modularity plays a significant role in the formation of patterns of activity on brain networks due to the small spectral gap of modular networks. As a byproduct, the modularity leads to the formation of macroscopic patterns, where the amount of activator and inhibitor on nodes in the same modules is approximately equal. These are similar to the patterns reported in the brain, and understanding the mathematical underpinnings of pattern formations can assist researchers in understanding brain development [1, 2] and neurological conditions [3].

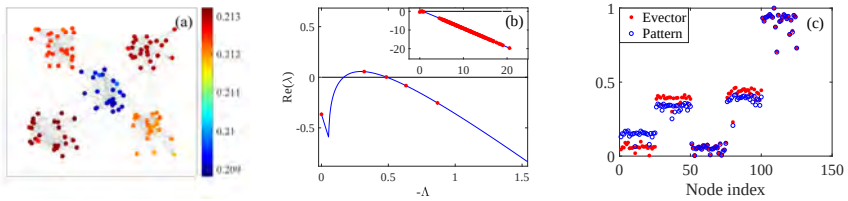


Fig. 1: (a) A macroscopic pattern on a modular network. (b) Dispersion relation of the pattern on the left. (c) Plotting the normalised eigenvector of the largest eigenvalue of the Jacobian, along with the normalised pattern. Note that the pattern and eigenvector show a similar structure.

References

1. Julyan H. E. Cartwright 11Labyrinthine Turing Pattern Formation in the Cerebral Cortex" J. theor. Biol. (2002) 217, 97103
2. Smith, G.B., Hein, B., Whitney, D.E., Fitzpatrick, D., Kaschube, M.: Distributed network interactions and their emergence in developing neocortex. Nat. Neurosci. 21, 16001608 (2018). doi:10.1038/s41593-018-0247-5
3. Demertzi, A. and Tagliazucchi, E. and Dehaene, S. and Deco, G. and Barttfeld, P. and Raimondo, F. and Martial, C. and Fernández-Espejo, D. and Rohaut, B. and Voss, H. U. and Schiff, N. D. and Owen, A. M. and Laureys, S. and Naccache, L. and Sitt, J. D. : Human consciousness is supported by dynamic complex patterns of brain signal coordination : Science Advances. 5 (2019). doi:10.1126/sciadv.aat7603
4. Atasoy, S., Donnelly, I., Pearson, J.: Human brain networks function in connectome-specific harmonic waves. Nat. Commun. 7, 110 (2016). doi:10.1038/ncomms10340
5. Hütt, M.T., Kaiser, M., Hilgetag, C.C.: Perspective: Network-guided pattern formation of neural dynamics. Philos. Trans. R. Soc. B Biol. Sci. 369, (2014). doi:10.1098/rstb.2013.0522
6. Turing, A.M.: The chemical basis of morphogenesis. Philos. Trans. R. Soc. London B. 52, 153197 (1952). doi:10.1007/BF02459572
7. Othmer, H.G., Scriven, L.E.: Instability and dynamic pattern in cellular networks. J. Theor. Biol. 32, 507537 (1971). doi:10.1016/0022-5193(71)90154-8
8. Nakao, H., Mikhailov, A.S.: Turing patterns in network-organized activator-inhibitor systems. Nat. Phys. 6, 544550 (2010). doi:10.1038/nphys1651
9. Asllani, M., Challenger, J.D., Pavone, F.S., Sacconi, L., Fanelli, D.: The theory of pattern formation on directed networks. Nat. Commun. 5, 19 (2014). doi:10.1038/ncomms5517
10. Asllani, M., Busiello, D.M., Carletti, T., Fanelli, D., Planchon, G.: Turing patterns in multiplex networks. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 90, 15 (2014). doi:10.1103/PhysRevE.90.042814
11. Bassett, D.S., Bullmore, E.: Small-world brain networks. Neuroscientist. 12, 512523 (2006). doi:10.1177/1073858406293182
12. Meunier, D., Lambiotte, R., Bullmore, E.T.: Modular and hierarchically modular organization of brain networks. Front. Neurosci. 4, 111 (2010). doi:10.3389/fnins.2010.00200
13. Sporns, O., Betzel, R.F.: Modular Brain Networks. Annu. Rev. Psychol. 67, 613640 (2016). doi:10.1146/annurev-psych-122414-033634

Functional Brain Network Topology Maps the Dysfunctional Substrate of Cognitive Processes in Schizophrenia

Rossana Mastrandrea[1], Fabrizio Piras[2] Andrea Gabrielli[3], Guido Caldarelli[1,3], Gianfranco Spalletta[2,4], and Tommaso Gili[1]

[1] IMT School for Advanced Studies, piazza S. Francesco, 19, 55100, Lucca, Italy
[2] IRCCS Fondazione Santa Lucia, Via Ardeatina 305, 00179 Rome, Italy
[3] Istituto dei Sistemi Complessi (ISC) - CNR, UoS Sapienza, Dipartimento di Fisica, Università "Sapienza"; P.le Aldo Moro 5, 00185 - Rome, Italy
[4] Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Tx, USA

1 Introduction

The investigation of brain functional organization, as obtained by resting state functional magnetic resonance imaging (rs-fMRI), has revealed differences in brain network topology in a number of psychiatric disorders, particularly in Schizophrenia [1].Brain intrinsic functional connectivity, captured by rs-fMRI [2], has shown alterations in specific brain circuits in schizophrenian [3], and evidenced the variability associated with this neuropsychiatric illness [4]; changes in global connectivity and alterations of local properties of the functional connectome have also been found [1, 4].

Recently network neuroscience [5, 6], the application of graph theory [7] to the study of brain networks, showed a widespread disturbances in the dynamics of large-scale brain networks [1, 8], and the alterations of the modular structure of the whole cerebral functional organization in schizophrenia [9]. Nonetheless, a unified description of the possible sources at the base of this mental illness is still under debate.

Specifically, the alterations in the global functional integration and the local functional connectedness of brain regions reported in literature, appear to be inconsistent across studies and limited to observations of a final effect.

In this paper, we analyzed the rs-fMRI of forty healthy subjects and fortyfour schizophrenic patients. We investigated the alteration of the hierarchical participation of brain regions to the whole network as a function of the correlation between their BOLD signals, used as edge-weights [10].We use percolation analysis, maximum spanning tree (MST) representation and allometric scales to study the backbone structure of the subjectwise functional brain network of healthy subjects and schizophrenic patients in order to assess the possible alterations induced by the illness.

2 Data

The BOLD signal of 40 healthy subjects and 44 Schizophrenic patients was registered for $6'$ (TR = 2"). After a detailed preprocessing of our data (control of time-lock cardiac and respiratory artifacts by means of linear regression, control for the effect of

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

low-frequency respiratory and heart rates, head-motion, slice timing corrections,spatial smoothin and normalization), we applied an AAL mask [11] to parcellate the human brain in 116 anatomical regions. We extracted the fMRI signals at voxel level, then we averaged them in each region of interest ending up with 116 BOLD time-series (180 time-points). We, then, computed their pairwise similarity using the Pearson's correlation coefficient obtaining a symmetric fully correlation matrix of dimensione $116x116$. We performed a subject-wise analysis due to the observed high level of heterogeneity of the chort of Schizophrenic patients. In order to not intoduce any arbitrary threshold, we performed a percolation a maximum spanning tree analysis looking at the whole matrix of squarred values. Indeed, the debate on the meaning of negative correlation values in fMRI studies is still open [12] without a general consensus militating in favour of their inclusion or exclusion in the analysis. We think they do not bring any relevant information to the scope of the actual study, therefore we will always refer to the squared correlation values focusing on the intensity of connection rather then its sign.

3 Results

Our findings demonstrate an augmented homogeneity of the weighted links distribution (correlation coefficient) in the connectivity intensity pattern of schizophrenic patients with respect to healthy individuals.

Looking at the percolation curves, schizophrenic functional networks appears more resistant to disconnection than healthy subjects one: the number of disconnected clusters at a given threshold is systematically greater in healthy subjects than in patients (fig.1 (a)). The weight distributions are very similar for the two groups of subjects, not explaining this difference (fig. 1 (b), top); on the contrary, an evident discrepancy can be found in the distribution of node degree computed for each percolated network (fig.1(b), bottom). This reveals a more homogenous distribution of intensity of connections in the human functional brain network of schizophrenic patients with respect to healthy subjects, sheding light on the difference in the giant component sizes.

The MST of the functional brain network in the two cases appear very similar, as confirmed by the allometric exponent used as a quantitative measure of the structural organization of the trees (fig.1(c)). Considering the widespread variation of the connectivity strength in schizophrenic patients, we introduce the concept of MST rank to deeply explore the topological properties of the MST. With *first rank* we refer to the usual MST computed on the functional brain network, the second (third) rank MST consists in the MST computed on the remaining network after the removal of all links corresponding to the first (second) rank MST and so on. Given a complete graph with N nodes it is always possible to decompose it in at most N/2 MSTs and to elicit the topological properties of trees characterized by weaker connections.

We find that, as the MST rank increases, the corresponding allometric exponent decreases both in schizophrenic and healthy subjects. However the rate of reduction in healthy subjects, leading to a net separation of the two groups at the third rank MST (fig. 1(d)-(f)).

The whole analysis (percolation, MST, MST ranks) suggests that (i) there is a reduced hierarchy and modular structure of the functional network due to a broader dis-

tribution of the connectivity weights; (ii) weaker and stronger links guarantee the same connection topology schizophrenia patients [1], (iii) there is a higher topological similarity of the MSTs of different rank, in schizophrenia patients suggesting that a given stimulus can engage a single functional connectivity path in healthy subjects, while it determines the simultaneous involvement of different ones in illness.

References

1. Bassett, D. S., Nelson, B. G., Mueller, B. A., Camchong, J. & Lim, K. O. Altered resting state complexity in schizophrenia. Neuroimage 59, 2196-2207 (2012).
2. Biswal, B. B. et al. Toward discovery science of human brain function. Proceedings of the National Academy of Sciences 107, 4734-4739 (2010).
3. Meyer-Lindenberg, A. et al. Evidence for abnormal cortical functional connectivity during working memory in schizophrenia. American Journal of Psychiatry 158, 1809-1817 (2001).
4. Lynall, M.-E. et al. Functional connectivity and brain networks in schizophrenia. Journal of Neuroscience 30, 9477-9487 (2010).
5. Bassett, D. S. & Sporns, O. Network neuroscience. Nature neuroscience 20, 353 (2017).
6. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10, 186-198 (2009).
7. Bollobas, B. Graph theory: an introductory course, vol. 63 (Springer Science & Business Media, 2012).
8. Uhlhaas, P. J. Dysconnectivity, large-scale networks and neuronal dynamics in schizophrenia. Current opinion in neurobiology 23, 283-290 (2013).
9. Bordier, C., Nicolini, C., Forcellini, G. & Bifone, A. Disrupted modular organization of primary sensory brain areas in schizophrenia. NeuroImage: Clinical 18, 682-693 (2018).
10. Mastrandrea, R., Gabrielli, A., Piras, F., Spalletta, G., Caldarelli, G., & Gili, T. (2017). Organization and hierarchy of the human functional brain network lead to a chain-like core. Scientific reports, 7(1), 4888.
11. Tzourio-Mazoyer, Nathalie, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15.1 (2002): 273-289.
12. Duff, Eugene P., et al. Disambiguating brain functional connectivity. NeuroImage 173 (2018): 540-550.

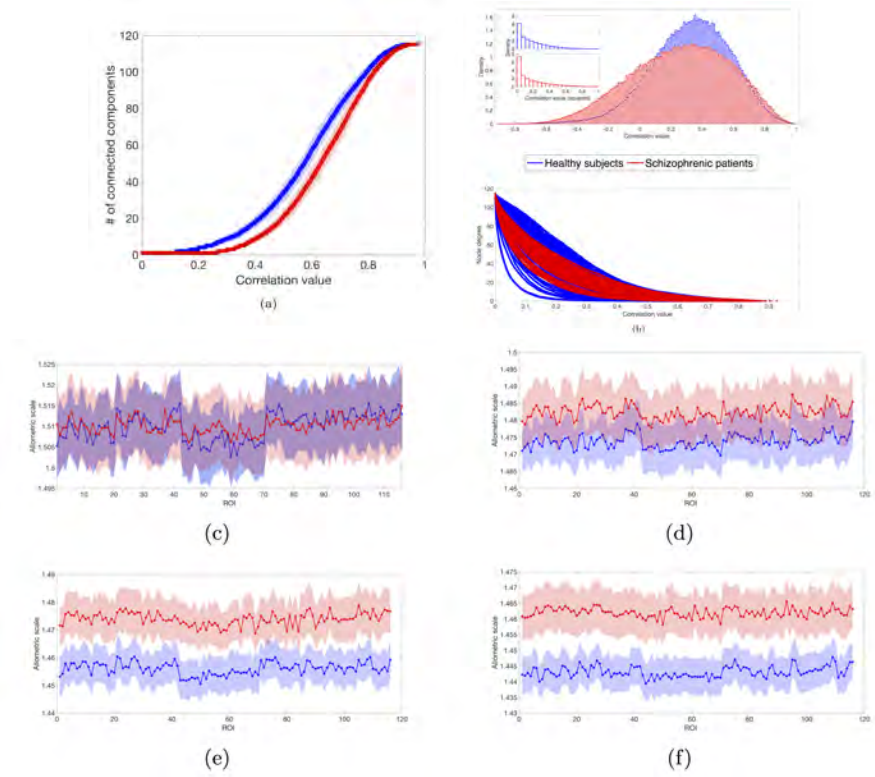


Fig. 1. (a) Percolation curves: number of connected components of the percolated network versus the related correlation threshold. The two curves represent the average of the individual percolation curves and are reported together with the 95% confidence interval; (b) Top: density of the distributions of correlation values of the human functional brain networks for all healthy subjects and schizophrenic patients pooled together in two distinct groups. Inset: distributions of the squared correlation values; Bottom: distribution of node degree computed in each percolated network and averaged over the total number of subjects in each group; (c)-(f) Allometric scale from the I to the IV MST rank computed for the MST of each healthy subject and Schizophrenic patient separately versus the related ROI-root. We reported the average allometric scale for the two groups together with their 95% confidence interval.

Paradigm filtering as a tool for new analysis of complex brain networks

Salvador Jimenéz[1], Jesús Tornero[2], Carlos Aguirre[3], and Laura Rotger[4]

[1] Dept. Matemática Aplicada a las TIC, ETSI Telecomunicación, Universidad Politécnica de Madrid, 28040-Madrid, Spain
s.jimenez@upm.es,
[2] Hospital los Madroños, M-501 Km 17, 900, 28690 Brunete, Madrid, Spain,
jesus.tornero@lmh.es,
https://hospitallosmadronos.es
[3] GNB, EPS, Universidad Autonóma de Madrid, 28049, Madrid, Spain
carlos.aguirre@uam.es
[4] Universidad Complutense de Madrid, 28040, Madrid, Spain

1 Introduction

We have known since the nineteenth century that the brain constitutes a huge and complicated structural network [1]. The latest advances in the study of complex systems have motivated new approaches and interpretations applied to brain structural and functional characterization [2].

Eguíluz et al. showed results of human brain functional networks with data extracted from temporal correlation between voxels using Pearson coefficients [3]. Two voxels were defined as functionally connected if their temporal correlation exceed a threshold value **rc**. Regardless the value of **rc** (from 0.5 up to 0.8) the brain functional networks from functional magnetic resonance imaging (fMRI) where characterized as free-scale networks [4].

In this work, functional brain networks are characterized through graph properties. Brain networks are built from functional magnetic resonance images taken at Hospital Los Madroños, Madrid, Spain. Then data is processed by a new set of two filters. The first filter computes the correlation between voxels and a type paradigm signal that represents the activation-deactivation of a finger-tapping task into the MRI equipment and then a second filter computes a correlation coefficient between voxels. In fig. 1 the activation zones, the paradigm signal and a voxel signal are plotted.

Once the network is created, the main topological graph measures are computed, mainly, degree distribution and its Shannon Entropy **S**, clustering coefficient **C**, characteristic path **L** and network efficiency **E**. See [5] for a description of these graph measures. Due to the huge size of the resulting graphs order ($\approx 2*10^5$) and size ($\approx 4*10^9$) a new C library for graph creation and computation has been built. This library supports different types of data structures (matrix, adjacency list, hash table and double linked list) for graph manipulation, selecting the optimal type of data for the measure to be computed. The library has been designed with the main objectives in mind to obtain a fast computing times with low memory consumption resources. The library (only C source code) can be downloaded from

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

https://www.dropbox.com/s/edqzua4bqn6apb9/small_graph.zip?dl=0

We will present results from 12-healthy volunteers under finger tapping task and blood-oxygen-level-dependent fMRI measurements [6]. Non-scale free networks with **rc** dependence were calculated in time and frequency domain.

2 Results

From the node degree distribution distributions and the parameters studied, we have observed behaviors that are repeated in all healthy subjects studied except in a pathological one.

In our implementation we have made two filters, one first with a paradigm that describes in a simplified way the tasks performed during the test, and a second filtering by correlations between each possible pair of voxels that pass the first filter. As we mentioned in the introduction, according to our knowledge it is the first time these two filters are applied together.

Observing the node degree distributions of all networks as well as the graph metrics studied (C, L, E and S) we have seen that the networks change with the threshold that we use for voxel-voxel filtering. We have observed the same characteristics both using motor paradigms as sensory. For low correlations, the networks resemble random graphs, while increasing the threshold, they become scale-free networks.

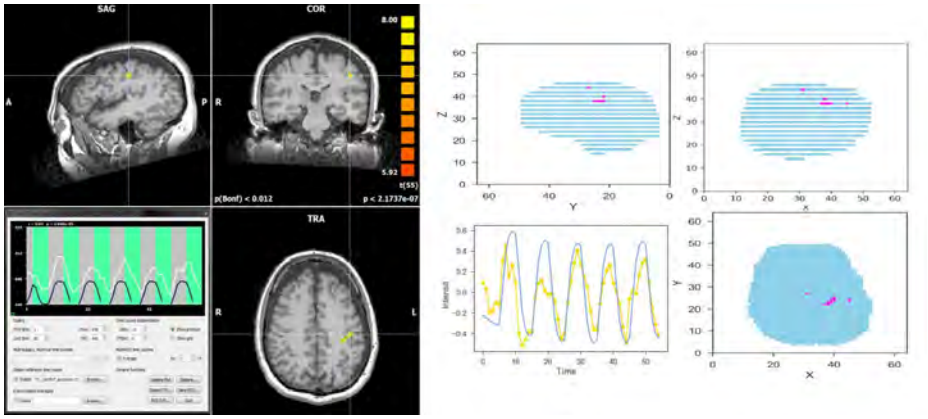


Fig. 1. Left: activation zones (yellow), for subject 3 and the left motor paradigm in the Sagittal, coronal and axial anatomical planes. Right: result obtained with our program Filtering. Below on the left is the paradigm signal in blue and the temporal trace of one of the selected voxels.

Summary. In this work, functional graph networks has been built using a new system of two filters, a first one comparing fMRI signals with a paradigm signal and then computing a correlation coefficient between voxels activity. Then the main topological

measures of the graph are computed by means of a new C library for graph computation capable of managing big networks with low computational time. Once the graph measures are done, these are compared for the case of normal patients and patients with some kind of disease such as ischemic strokes.

3 Acknowledgments

CA is supported by MINECO/FEDER PGC2018-095895-B-I00

References

1. Swanson, L.W.: Brain architecture. Oxford University Press. Oxford (2005)
2. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature reviews neuroscience 10, 186–198 (2009)
3. Eguíluz, V.M., Chialvo, D.R., Cecchi, G.A., Baliki, M., Apkarian, A.V.: Scale free brain functional networks. Physical Review Letters 94, 018102-1-018102-4 (2005)
4. Benesyt, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. Springer. Berlin-Heidelberg. 1–4 (2009)
5. Rubinov, M., Sporns O.: Complex network measures of brain connectivity: Uses and interpretations. Neuroimage 52, 10591069 (2010)
6. Filippi, M. (Ed): fMRI techniques and protocols. Springer-Humana Press. New York (2016)

Multilayer brain networks with time-evolving nodes and analyzing network motifs in them

Tarmo Nurmi[1], Onerva Korhonen[2], and Mikko Kivelä[1]

[1] Aalto University, Department of Computer Science, Finland,
[2] Université de Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, Lille, France

Human brain function can be represented as a network of brain regions, Regions of Interest (ROIs). ROIs form the nodes of the network and edges represent correlation of brain activity between ROIs. In functional brain networks constructed from functional magnetic resonance imaging (fMRI) data, ROIs consist of several measurement volume elements, voxels. ROIs are assumed to be functionally homogeneous or, in other words, each voxel in a ROI is assumed to behave similarly in time. These networks have long been analyzed as static, i.e. unchanging in time, despite changes in brain activity over time [1]. Additionally, ROIs have been seen as static regions between which edge patterns may change in time but not the shapes and sizes of the ROIs themselves. According to recent studies, this view may be inaccurate: in various brain parcellations (divisions of the brain into ROIs), the functional homogeneity of ROIs is low and varies in time [3, 6]. In other words, voxels in the same ROI show different dynamics and are therefore hardly more correlated than voxels in different ROIs (see Fig. 1). Evidently, there is a need for defining brain network nodes in a data-driven way that maximizes ROIs' homogeneity and allows their shape, size, and location to change over time.

The multilayer network model of brain function. We incorporated the flexible definition of ROIs in different time intervals, the connections between ROIs within each time interval, and the similarity information of ROIs in different time intervals into one multilayer network [2]. In this network, time intervals correspond to layers and ROIs (nodes) within each layer can be defined independently of other layers, resulting

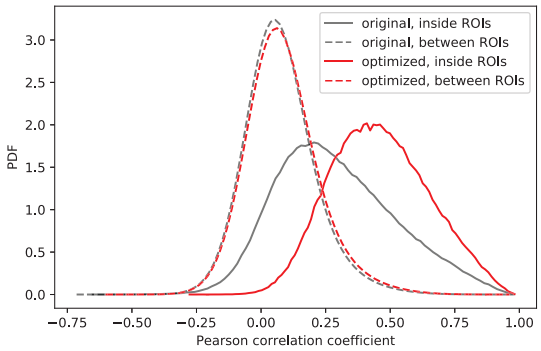


Fig. 1. Distributions of Pearson correlation coefficients between voxel time series within and between ROIs (original = the static Brainnetome parcellation).

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

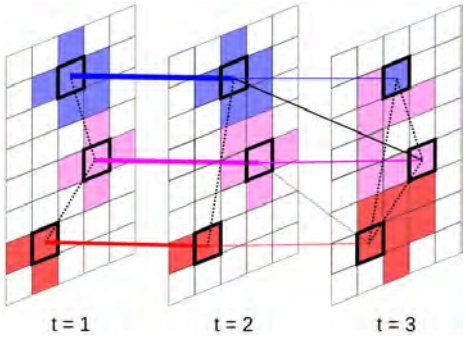


Fig. 2. Schematic illustration of the functional temporal multilayer brain network. The three layers (t = 1, 2, and 3) correspond to three time intervals in consecutive order. The background grid corresponds to voxels, and ROIs are defined as contiguous sets of voxels (red, blue, and pink areas) clustered using data from that specific time interval. The ROIs within each layer are connected by edges representing functional similarity between them. Interlayer edges are weighted according to the Jaccard index of the sets of voxels in ROIs on consecutive layers. Not all voxels are necessarily included in the ROIs.

in ROIs that change in time. The multilayer network is constructed as follows: Within each layer, voxels are clustered into ROIs according to functional homogeneity while keeping the ROIs spatially contiguous. The increased functional homogeneity of optimized ROIs leads to markedly higher correlations between voxel time series within ROIs than between ROIs (Fig. 1). Inside each layer, the networks are constructed as is typical in the literature for static brain networks such that ROIs are connected with edges weighted according to the correlation between ROI time series. Then, the ROIs on a specific layer are connected to ROIs on the temporally preceding and following layer with edges weighted according to the magnitude of overlap of the sets of voxels constituting the ROIs, measured e.g. by the Jaccard index $J(S_1,S_2) = |S_1 \cap S_2| / |S_1 \cup S_2|$, where S_1 and S_2 are the sets of voxels in the two ROIs on consecutive layers. Fig. 2 illustrates the resulting multilayer network, where the ROIs (red, blue, and pink areas) change between layers.

Multilayer network motifs. To understand the function of these temporal multilayer networks, we disassembled them into smaller building blocks known as graphlets and motifs [4, 5]. Motifs are defined as isomorphic equivalence classes (isomorphism classes) of subnetworks that are statistically more or less prevalent in one data set than in another or in a mathematical null model. Therefore, they can be used to find differences in brain function between groups of subjects or to analyze the fundamentals of human brain function with respect to a baseline model. Multilayer motifs represent patterns of network connectivity both within each layer and between layers, making them an ideal tool for capturing both the interplay of brain regions within layers and the changes in brain regions from one layer to the next. Since motif analysis requires no mapping of nodes between different subjects, networks with subject-specific brain regions can be properly compared. Despite the usefulness of motifs in analyzing the structure of single-layer

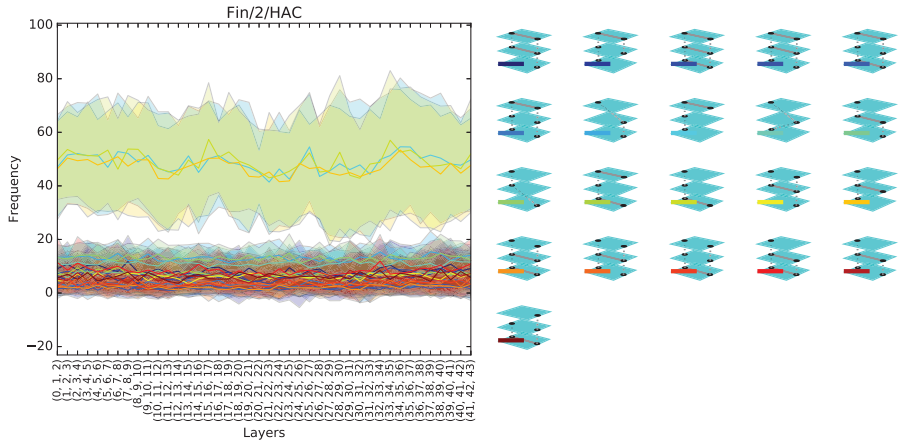


Fig. 3. Time series of the number of isomorphism classes of specific size in multilayer networks constructed from fMRI images. Each time point is a collection of consecutive layers, and the horizontal (time) axis advances in steps of one layer. Each isomorphism class has its own color, shown on the right. The shaded areas show the standard deviation between subjects (N = 25).

networks [4], tools for analyzing multilayer network motifs have been largely absent before this work.

Proof-of-concept. We constructed temporal multilayer brain networks from an fMRI data set collected at Aalto Magnetic Imaging (AMI) Center and enumerated the subnetworks in them. Fig. 3 shows isomorphism class appearance frequencies with respect to time in a set of subjects for multilayer isomorphism classes of specific size. We compared these time series between subject groups to discern differences in brain function, and compared the frequencies to ones obtained for a random null model to find motifs related to brain function in general. The framework enables a more accurate network representation of time-evolving brain activity and offers new perspectives in the study of fundamental brain function as well as opportunities for medical applications.

References

1. Bassett, D.S., Sporns, O.: Network neuroscience. Nature Neuroscience 20(3), 353 (2017)
2. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. Journal of complex networks 2(3), 203–271 (2014)
3. Korhonen, O., Saarimäki, H., Glerean, E., Sams, M., Saramäki, J.: Consistency of regions of interest as nodes of fmri functional brain networks. Network Neuroscience 1(3), 254–274 (2017)
4. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
5. Pržulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? Bioinformatics 20(18), 3508–3515 (2004)
6. Ryyppö, E., Glerean, E., Brattico, E., Saramäki, J., Korhonen, O.: Regions of interest as nodes of dynamic functional brain networks. Network Neuroscience 2(4), 513–535 (2018)

Dynamic functional connectivity through graph metrics for classification in motor imagery BCIs

Paula Rodrigues[1,2], Arnaldo Fim Neto[1,2] André Takahata[1,2], and Diogo Soriano[1,2]

[1] Federal University of ABC, Engineering, Modeling and Applied Social Sciences Center, São Bernardo do Campo, Brazil
[2] Brazilian Institute of Neuroscience and Neurotechnology (BRAINN)

1 Introduction

The scalp electroencephalogram (EEG) allows one to capture the brain electrical activity under different mental tasks or event-related stimuli [1] and defines a powerful brain investigation paradigm. The EEG observations can also reveal information related to the brain activity organization, which implies in quantifying how this activity is spatially (anti-) correlated in such mental tasks by means of the functional connectivity evaluation [2]. Besides, the interconnectivity of the brain regions can be modeled as a complex network and all the framework associated with a graph-based approach can be applied to elucidate neural mechanisms in healthy and abnormal brain [2].

Recently, it has been shown that graph-based metrics regarding EEG functional connectivity can distinguish, with a promising performance, motor imagery (MI) tasks (left- and right-hand MI), aiming the design of more robust brain-computer interfaces (BCIs), i.e. systems which aim to translate information from central nervous system activity into output commands [3]. However, up to now, the dynamics of the MI process and its non-stationary behavior have not been fairly considered. Some recent works also suggested that the dynamics of functional connectivity can provide useful information for movement x rest discrimination [4]. In this context, this work presents a classification performance analysis considering the eigenvector centrality and strength metrics in a 52 subjects dataset [5], seeking to find out whether the graph metrics would provide high performance to distinguish the left- and right-hand mental tasks regarding the dynamic functional connectivity considering a template matching classification approach. This work may be relevant for designing new strategies for MI-based BCIs, unveiling new biomarkers to distinguish mental tasks.

2 Material and Methods

A BCI dataset comprising two imagery tasks (left and right hand) for 52 subjects was used in this analysis. Signals were recorded during 3 s with 64 Ag/AgCl electrodes sampled at 512 Hz. For each task, 100 trials were available. More information about the dataset can be found in [5]. During offline preprocessing, signals were filtered spatially with Common Average Reference and, then, bandpass filtered (8-15 Hz) in which the mu band, EEG frequencies related to the motor cortex, was conserved [1, 6].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The dynamic functional connectivity was computed by using a 40-points sliding window without overlap. Regarding only the 21 motor cortex electrodes (FC5, FC3, FC1, C1, C3, C5, CP5, CP3, CP1, CPZ, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, CP6, CP4, CP2 – international 10-10 system), the connectivity was estimated by means of the Pearson correlation coefficient. Only the positive correlation values were considered, defining a weighted connectivity matrix for each window.

In this work, eigenvector centrality (EC) and strength (ST) were chosen to characterize the dynamic connectivity among electrodes, which correspond to the nodes of our graph. Strength is defined as the sum of the edge weights of a node, similar to the node degree for unweighted graphs. On the other hand, EC metric considers not only the connections of a node but also the importance of its neighbours in the network [2]. For comparison, bandpower (BP) was computed through the same sliding window, defining a dynamic bandpower. In this analysis, only the connectivity of the C3 and C4 electrodes were considered, in agreement with previous work [6].

For the classifier training purpose, time-courses templates matching for each mental imagery task were obtained for C3 and C4 by computing the median of the graph metrics estimated for each window throughout the training trials. The median time course was then smoothed by a moving average at 10 points to capture the main behavior. In order to evaluate the classification performance, a 10-fold cross-validation was used. Lastly, classification was estimated based on template matching by means of the minimum mean square error between a given smoothed time course of the trial (left or right hand) and the respective (right and left hand) templates of a specific electrode.

The accuracy was presented as mean $\pm$ standard deviation. At least 3 statistical tests were carried out in order to confirm the non-normality of the data. The Friedman test was applied to evaluate the statistical difference among the groups EC, ST and BP as well as the post-hoc analysis using Dunn's multiple comparison test to provide difference for each paired combination of the groups.

3 Results and Conclusion

Figure 1 presents the accuracy of the approaches considering the 52 subjects in the dataset for C3, C4 and their combination. Analysing the accuracy obtained by the graph metrics as well as bandpower for C3 electrode, it was observed significant statistical difference (p-value < 0.01, Friedman test, N = 52) among the mean of the groups EC (0.56 ± 0.08), ST (0.55 ± 0.08) and BP (0.52 ± 0.05). Post-hoc analysis unveiled that the accuracy of the groups EC and ST ($p < 0.05$) and EC and BP ($p < 0.01$) were significantly different.

On the other hand, by analysing the C4 electrode, it was observed statistical difference ($p < 0.001$, Friedman test, N = 52) among the group and the mean of EC (0.57 ± 0.07) and BP (0.54 ± 0.07) were different ($p < 0.0001$), but ST (0.56 ± 0.06) was distinct from BP ($p < 0.05$) and not from EC as previously observed for C3 electrode. In addition, analysing the combination of C3 and C4 electrodes, although the mean of EC (0.57 ± 0.08), ST (0.58 ± 0.09) and BP (0.56 ± 0.10) were statistically different ($p < 0.05$, Friedman test), no difference was found when analysed each post-hoc combination of the groups.

From the results, even though EC and ST were not statistically different from BP when both electrodes were considered, they showed promising results considering the individual electrodes, specially the EC metric, despite the lower accuracy for practical BCI applications, in which accuracies higher than 0.80 are typically expected. More study is required in order to improve the obtained results, but this exploratory work encourages further evaluation of the dynamic functional connectivity through graph metrics in MI-BCIs.

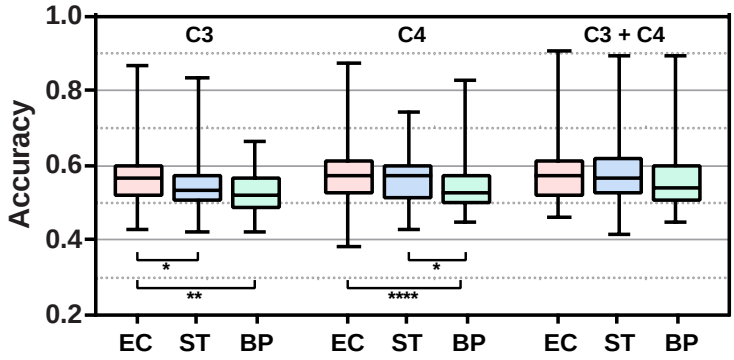


Fig. 1. Boxplot considering the accuracy of the 52 subjects for eigenvector centrality (EC), strength (ST) and bandpower (BP) for C3, C4 and their combination. * $p < 0.05$; ** $p < 0.01$; **** $p < 0.0001$

4 Acknowledgement

The authors thank the financial support from CNPq 305616/2016-1, FAPESP 2013/07559-3, 2019/09512-0, FINEP 01.16.0067.00, CAPES 2019/1814368 and UFABC.

References

1. Wolpaw, J., Wolpaw, E.W.: Brain-Computer Interfaces: Principles and Practice. Oxford University Press (jan 2012)
2. Sporns, O.: Networks of the Brain. The MIT Press (2011)
3. Rodrigues, P.G., Filho, C.A.S., Attux, R., Castellano, G., Soriano, D.C.: Space-time recurrences for functional connectivity evaluation and feature extraction in motor imagery brain-computer interfaces. Medical & Biological Engineering & Computing **57**(8) (may 2019)
4. Williams, N.J., Daly, I., Nasuto, S.J.: Markov model-based method to analyse time-varying networks in EEG task-related data. Frontiers in Computational Neuroscience **12** (September 2018)
5. Cho, H., Ahn, M., Ahn, S., Kwon, M., Jun, S.C.: EEG datasets for motor imagery brain-computer interface. GigaScience **6**(7) (2017) 1–8
6. Li, F., Peng, W., Jiang, Y., Song, L., Liao, Y., Yi, C., Zhang, L., Si, Y., Zhang, T., Wang, F., Zhang, R., Tian, Y., Zhang, Y., Yao, D., Xu, P.: The dynamic brain networks of motor imagery: Time-varying causality analysis of scalp EEG. International Journal of Neural Systems **29**(01) (January 2019) 1850016

Interface to Functional Connectivity Analysis of EEG Signals using Complex Networks

Daniel M. Alves, Paula G. Rodrigues, André K. Takahata, and Diogo C. Soriano

ABC Federal University (UFABC), Center for Engineering, Modeling and Applied Social Sciences (CECS), Santo André CEP 09210-580, Brazil,
martins.daniel@ufabc.edu.br

1 Introduction

Brain-computer interface (BCI) systems allow brain activity to control computers or external devices, with the primary purpose of providing communication capabilities to severely incapacitated people by neurological neuromuscular disorders, such as amyotrophic lateral sclerosis, stroke, or spinal cord injury [1]. Electroencephalography (EEG) is one of the most popular techniques to record brain signals using electrodes (receptors) placed on the scalp. Compared to other techniques, their main advantages are low cost, relative ease of use and excellent time resolution (milliseconds) [1].

A complex network can be represented by a graph that consists in a large collection of nodes connected by edges that can represent the associations of any quantity: people, computers, biological cells, etc [2]. In a graph there are several measures that can be generated to analyze complex networks, for example: degree, clustering coefficient, eigenvector centrality and betweenness centrality [3].

Considering the spatiotemporal distribution of brain activities, the human brain is like a complex network formed by a large number of connected cortical regions. Information related to, for example, motion imagery is constantly integrated and processed between scattered and specialized brain regions [4]. Thus, to understand brain functioning in a given action, it is necessary to study not only isolated regions, but also the connectivity between them. There are three brain connectivity types defined in the literature: Structural, which refers to a set of physical or structural connections between neural elements, functional, which captures deviation patterns from statistical independence between distributed and often spatially remote neuronal units, and effective, similar to functional but describes the causal effects network between neural elements [4].

Keeping these concepts in mind, this paper propose a user-friendly interface to analyze functional connectivity from temporal observations (e.g. EEG) and their respective network measures (degree, clustering coefficient, centralities, etc.). In this sense, it seeks here the integration of the interface with the data transfer system of the OpenBCI platform, which is intended for experiments involving EEG, in particular, the development of brain-machine interfaces [5]. The interface was developed in JAVA language and is based on tools already available in the GraphStream toolbox [6], developed in the same language.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

Two structures were developed, a NodeJS-based web service [7] that get signals from OpenBCI EEG headset and a Java-based interface for getting data from text files or data buffer in real-time. Fig. 1 shows the integration between them and OpenBCI platform.

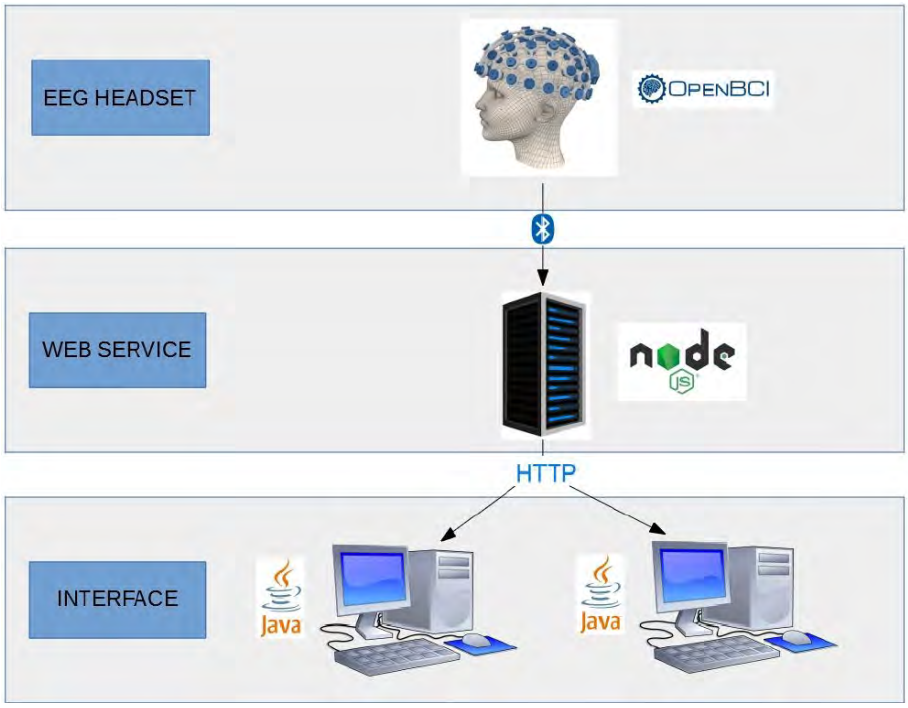


Fig. 1. Structures and their interactions

The developed NodeJS-based web service aims to communicate with the OpenBCI platform by bluetooth and make the EEG signals collected available in network through HTTP protocol and using JSON format. This way any application connected to this network can access the data, enabling local and remote processing.

The developed interface uses networked signals to turn them into complex networks to analyze functional brain connectivity. It allows users to set: a) the use of common average reference (CAR) spatial filter; (b) the updating graph rate, i.e., data samples for each time window; c) the number of overlapping samples between consecutive windows; d) Pearsons threshold correlation between the electrodes to define functional connectivity. Four graph measures can be evaluated by the interface using the GraphStream toolbox: node degree, clustering coefficient, eigenvector centrality and betweenness centrality. The results with the desired metrics can be exported to a text file.

Fig. 2 show the Java-based interface parts. On the left you can see settings options commented above, while on the upper right is the graph generated by the interface. The

bottom right displays a line graph showing the evolution of the chosen measure over time.

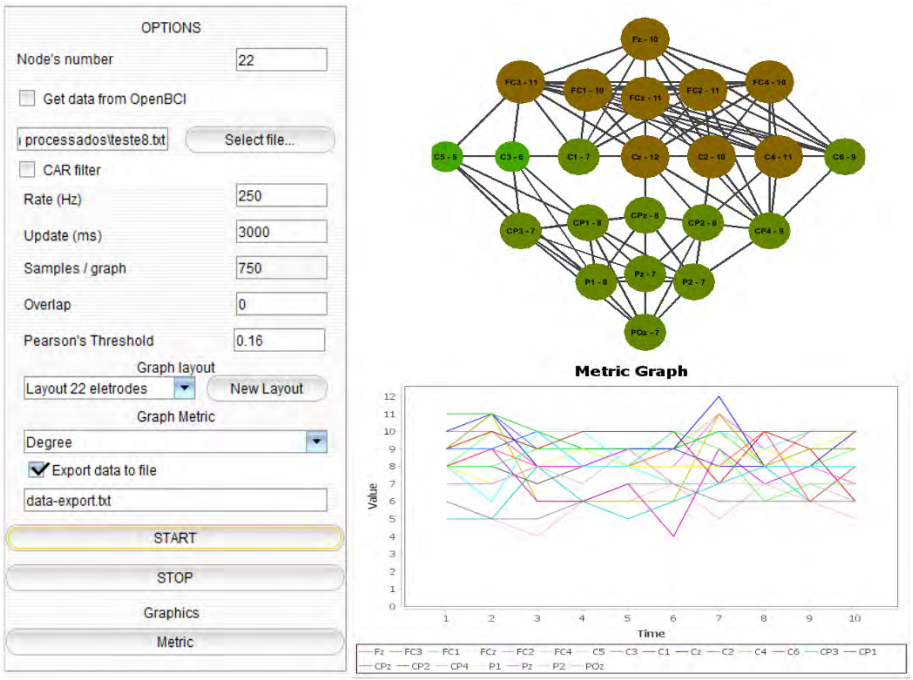


Fig. 2. Java-based interface parts

Summary. This work presents a graphical interface for functional connectivity analysis and brain functional organization investigation. The interface was integrated with low cost Open BCI acquisition hardware, using a NodeJS-based web service that make the EEG signals collected available in network, and GraphStream toolbox, a computational efficient graph analysis environment with high potential application for online operation, which outlines a natural perspective of this work.

References

1. Hassanien, A. E., Azar, A. T.: Brain-Computer Interfaces. Cham: Springer International Publishing. E 74 (2015)
2. Steen, M.V.:Graph Theory and Complex Networks: An introduction. 1(3) (2010)
3. Junker, B. H., Schreiber, F.: Analysis of biologicals networks. John Wiley & Sons, Inc. (2008)
4. Sporns, O.: Networks of the Brain. The MIT Press E 1542 (2015)
5. OpenBCI: https://openbci.com (2019)
6. GraphStream Project. http://graphstream-project.org (2016)
7. Node.js: https://nodejs.org (2019)

A multiplex approach to neuroimaging: applications to Alzheimer, Parkinson and aging

Nicola Amoroso[1,2], Loredana Bellantuono[1], Alfonso Monaco[2], Sabina Tangaro[2], and Roberto Bellotti[1,2]

[1] Università degli studi di Bari "Aldo Moro", Dipartimento Interateneo di Fisica, Bari, Italy, loredana.bellantuono@ba.infn.it,
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy

1 Introduction

Magnetic Resonance Imaging (MRI) brain scans are effectively used for early diagnosis of neurodegenerative diseases. Such pathologies affect the human brain with heterogeneous and complex patterns, but it is possible to identify some common traits; above all brain atrophy, i.e. the loss of neurons and the consequent increment of cerebrospinal fluid, is a specific feature typically outlined by MRI scans. Since modifications in the brain structure can be extremely subtle to detect in early stages of the disease, neuroradiologists have to perform extremely time consuming investigations to explore MRI data in detail, with possible subjectivity issues. Due to the increasing availability of huge neuroimaging databases, several studies have recently explored the possibility to design and implement machine learning approaches, in which sophisticated algorithms try to learn patterns related to disease and provide diagnostic decision support systems. In recent years, the feasibility of these approaches has been investigated with mixed fortunes; international challenges launched to compare different algorithms with common validation strategies and data have highlighted benefits and limits of these methods [1], eventually suggesting the need for a more robust pipeline of analysis. In order to obtain structural features (volumes, surfaces, curvatures, ...) from MRI data and feed machine learning algorithms, cumbersome pre-processing steps are required. Specifically, brain scans are typically registered in order to provide a common reference space for all subjects and are then segmented to focus on anatomical districts of interest, together with their quantitative features. However, this base of knowledge can be dramatically affected by registration errors and uncertainties, which prevent the possibility to get accurate results.

Promising strategies to overcome this issue are provided by Network Physiology, an emerging framework which relates organs' states and functions to the collective dynamics of their constituents, using modeling techniques based on complex networks [2–5]. In particular, this approach allows to examine the time-dependent interactions among brain districts [6], opening up the possibility to unveil the impact of their dynamics on the occurrence of anomalous aging and neurodegenerative diseases. In order to acquire new insights into biomarkers of these conditions, we propose a new method within Network Physiology, based on multiplex for early diagnosis of Alzheimers and Parkinson's diseases [7, 8] and predictions on brain aging [9].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

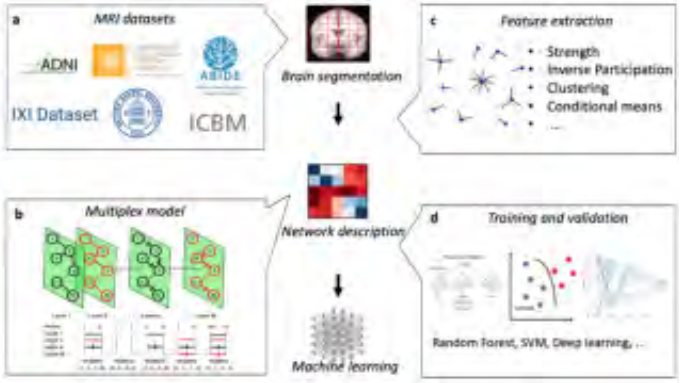


Fig. 1. A schematic overview of the proposed framework of analysis. Starting from structural MRI data (a), we achieve a multiplex description (b) and use this model to compute mathematical features characterizing the nodes of the network (c); finally, these features are used to feed supervised learning algorithms, such as Random Forest, Support Vector Machines or Deep Neural Networks, which perform classification and regression tasks (d).

2 Methods

In our approach, MRI brain scans of different subjects are registered to the common MNI152 template. Then, every scan is segmented into a three-dimensional grid of rectangular boxes, called *patches*, each consisting of a fixed number of voxels. Patches represent the nodes of an undirected subject-specific complex network, in which the link connecting a given pair of patches is weighted by the Pearson's correlation between their respective gray level distributions. Hence, link weights yield a quantitative measure of morphological similarity between brain districts. Since the spatial distributions of white and gray matter and cerebrospinal fluid in the brain change with age and pathological conditions, we model the whole cohort as a multiplex, with subject-specific networks as layers. We compute a set of nodal metrics (strength and inverse participation, together with their conditional means at fixed node degree) for each layer. Then, to capture inter-subject variation, we construct the aggregate adjacency matrix of the cohort and compute the multiplex versions of the aforementioned nodal metrics by weighting each node contribution with its aggregate degree. The single-layer and multiplex metrics thus extracted for each subject are employed as features to train Machine Learning (ML) or Deep Learning (DL) algorithms, performing the following tasks:

- classification, to distinguish subjects affected by Alzheimer's disease (AD), mild cognitive impairment (MCI) or Parkinson's disease (PD) from healthy controls (NC);
- regression, to predict the subjects' age, which is a continuous variable.

Cross-validation techniques are adopted to test the developed algorithms on the whole dataset and ensure reliability in performance evaluation. A schematic overview of the described pipeline is displayed in Fig. 1. The proposed approach requires only a rough spatial overlap for the scan of different individuals and does not require any *a priori*

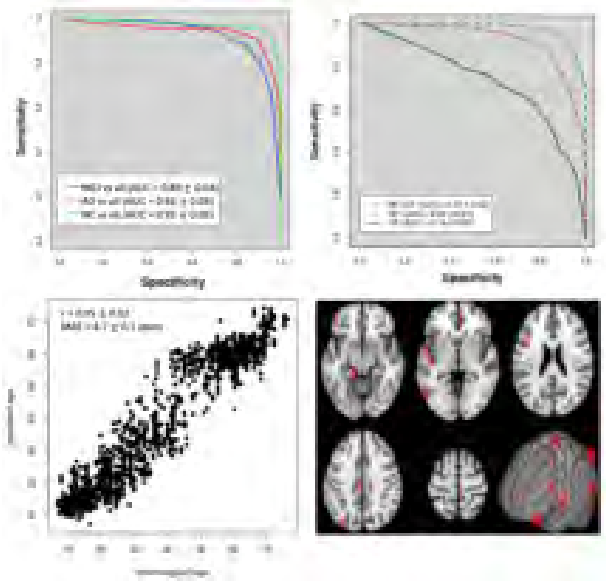


Fig. 2. Top left. Classification performance for Alzheimer's disease in terms of area under the receiver operating characteristic curve (AUC): patients (AD), healthy controls (NC) and subjects affected by mild cognitive impairment (MCI) are enrolled. **Top right.** Classification performance for Parkinson's disease in terms of AUC: network features (NF) and clinical features (CF) classification performances are compared. **Bottom left.** Regression performance for brain aging: scatterplot denoting predicted and chronological age of the subjects, together with Pearson correlation coefficient (r) and Mean Absolute Error (MAE) of the model. **Bottom right.** The methodology also allows to rank the statistically significant features and localize them in the brain (red regions); the reported case refers to aging.

segmentation, which ensures robustness against errors in registration and segmentation steps.

3 Results

The pipeline described in Section 2 is repeated by varying the patch size, the employed algorithms and their tuning parameters, in order to determine the best configurations for each of the considered classification and regression tasks. We demonstrate for all studies that an optimal scale exists in relation to the particular disease: performances are maximized by parceling the brain scans in patches with volume 3000 mm^3 for AD and aging, and 125 mm^3 for PD. In all our classification models, nodal metrics undergo a feature selection process, operated by a Random Forest wrapper. For AD and MCI, the proper classification task is performed by a further Random Forest block, with results displayed in the top left panel of Fig. 2. In the case of PD, instead, we compare the performances of Support Vector Machine, Random Forest, Naive Bayes and Neural Network classifiers, after optimizing them via parameter tuning. The multiplex model of brain connectivity yields a robust diagnosis independent of the choice of the classifier,

although Support Vector Machine reaches slightly better results. As shown in the top right panel of Fig. 2, the use of network features (NF) in PD identification algorithms entails a relevant improvement of classification accuracies with respect to the ones obtained from just the clinical features (CF). In order to assess the validity of our classification algorithms, we compare their performances with those of standard approaches, in which feature extraction from multiplex is replaced by Voxel Based Morphometry and Region-of-Interest methods [10, 11]; in all the examined cases, the network approach provides more accurate results. For the regression task of brain age prediction, we considered DL and ML approaches (Ridge and Lasso regression, Random Forest and Support Vector Machine), finding their optimal configurations through systematic exploration of the respective parameter spaces. The DL regression algorithm, whose results are shown in the bottom left panel of Fig. 2, yields better performances than the ML methods. Moreover, the accuracies of our multiplex model compare favorably with state-of-the-art techniques [12, 13]. The proposed framework enables us to implement regression algorithms that provide accurate information on brain aging and allow to rank the structural features based on their statistical significance, associating them to a specific cerebral patch. Bottom right panel of Fig. 2 shows brain districts, highlighted in red, that exhibit the most relevant structural changes in aging processes.

References

1. Bron, E.E., et al.: Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562-579 (2015).
2. Bashan A., et. al.: Network physiology reveals relations between network topology and physiologic function. Nature Communications 3, 702 (2012).
3. Bartsch R.P., et. al.: Network Physiology: how organ systems dynamically interact. PLOS ONE 10(11), e0142143 (2015).
4. Ivanov P.C., et al.: Focus on the emerging new fields of network physiology and network medicine. New Journal of Physics 18(10), 100201 (2016).
5. Ivanov P.C. and Bartsch R.P.: Network Physiology: Mapping Interactions Between Networks of Physiologic Networks. In: D'Agostino G., Scala A. (eds) Networks of Networks: The Last Frontier of Complexity. Understanding Complex Systems. Springer, Cham, 203 (2014).
6. Liu K.K.L., et. al.: Plasticity of brain wave network interactions and evolution across physiologic states. Frontiers in Neural Circuits 9, 62 (2015).
7. Amoroso, N., et al.: Multiplex Networks for Early Diagnosis of Alzheimer's Disease. Frontiers in Aging Neuroscience 10, 365 (2018).
8. Amoroso, N., et al.: Complex networks reveal early MRI markers of Parkinsons disease. Medical Image Analysis 48, 12-24 (2018).
9. Amoroso, N., et al.: Deep Learning and Multiplex Networks for Accurate Modeling of Brain Age. Frontiers in Aging Neuroscience 11, 115 (2019).
10. Ashburner, J. and Friston, K.J.: Voxel-based morphometry–the methods. NeuroImage 11(6), 805-821 (2000).
11. Fischl, B.: Freesurfer. NeuroImage 62(2), 774-781 (2012).
12. Cole, J.H., et al.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115-124 (2017).
13. Franke, K., et al: Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. NeuroImage 63(3), 1305-1312 (2012).

Part XIV

Networks in Finance and Economics

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The Firm as an Emergent Phenomenon

Dirk Bruin*

Vrije Universiteit Brussel - Center Leo Apostel (CLEA)
Rue de la Stratégie, Brussels, Belgium // www.ecco.vub.ac.be
dirkpbruin@gmail.com // www.magrathea-tlc.nl

1 Introduction

A Theory of the Firm is an economic theory to explain the nature of the firm, including its existence, its behavior and structure, and its relationship to the markets. A question reflecting the traditional economist's view on the existence of firms, is: Why should human behavior cohere into business organizations if markets are in place to coordinate the supply of goods and services? The answer provided by transaction cost theory is that the entrepreneur assesses the differential of the transaction cost internal to the firm and external (a market). When that transaction cost of a business activity is lower 'inside' the firm, the entrepreneur is predicted to include it and vice versa [Coase 1937, Williamson 1979]. Note that this argument assumes that the smallest component of markets and firms is a person.

However, this assumption denies the fact that the influence of an individual person on the behavior of the firm is limited. Evidence for this is provided by the argument that the impact on the behavior of a firm (as such) of a person tends to be limited. Also it is more plausible that firms are adopted by a niche than that they adapt into a niche of choice [Alchian 1950]. Thirdly, recent upheaval in the banking, pharmaceutical, and automotive industry sectors has proven them persistent to desired change [cf. Luyendijk 2005]. Lastly evidence shows that the correlation between individual personal behavior (performance) and firm behavior is weak [CBS 2002, Kahneman 2005]. Current theory of the firm however assumes a strong influence. A central notion in this hypothesis is that the idea is separated from the person.

My research concerns the development of a theory to explain the nature of the firm by answering the question: What is a firm? anew, and, more specifically: How is it conceived, How does it cease to exist, and What are its structure and behavior? The final objective is to better explain firms' behavior vis-a-vis the people associated.

2 Preliminary Considerations

Let us define ideas as answers to questions. This theory is based on the notion that a variable yet persistent pattern of behavior arises from an evolving

* I thank Prof Heylighen, Dr Lenartowicz and Dr Velóz-Gonzales for their contributions to the development of this theory under construction.

coherent system of ideas. More precisely: the ideas in conjunction with the behavior they induce evolves, instead of the traditional Platonic view, of a firm as a mathematical dot related to other such objects situated in markets and pursuing equilibrium. For example a practical organization of people is often assumed to an aberration from an ideal organizational type and deviations are assumed to be manageable to be directed to reach the ideal [Mintzberg 1983 i.a.]. This present hypothesis assumes that the entire firm including its organization consists of complexes of ideas and coherent behavior of people in the sense of behavioral patterns instead of ideal objects gravitating towards their ideal.
This implies that the ideas themselves are not essential and symbolised denotatively but proscriptive and Just-So [Kipling 1902]. Theory development is therefore founded on process ontology: differences and processes instead of objects and their relations are at the core. In a philosophical perspective: the real is developed from restriction of the virtual in the actual to produce a new actual, etc.[cf. Deleuze 1968]. This takes place in processes of individuation, where the structure of a system leads to its operations and vice versa, and transduction between the system and its neighbors [cf. Simondon 1958, Deleuze 1968].

3 Theoretical Framework

This theory builds on these concepts and theories: explanatory coherence theory is applied to ideas to explain how ideas connect to others to form maximally coherent complexes [cf. Thagard 2007]. Micro-sociology [cf. Knorr-Cetina 1988] and enactment [Weick 1988] explain how ideas lead to behavior through acting on what is not there. An abstract view is taken on theoretical ecology (autopoiesis) to explain how social systems self-replicate including their functions to self-replicate [cf. Maturana, Varela 1972]. Invoking this theory requires that discussions concerning the transposition from the biological to the cultural be dealt with. One important bridging element is the distinction of the person biologically evolving to a internal social entity versus the evolution of external social entities and forms: subjects of evolution but different species [Lenartowicz 2016, Varela 1997].
Social systems theory and importantly the concept of double contingency explain how events of communication come to be self-replicating autopoietic social systems [Luhmann 1996, Varela 1997, Varela, Thompson, Rosch 1992]. This is the basis for the explanation of coherence of human behavior as the behavioral pattern which make sense and can be made sense of. A body of widespread ideas originating from Western culture is conjugated with this hypothesis, namely those encompassed by The Market System and Belief in Progress [Goudzwaard 1982, Galbraith 2004]. These ideas are deconstructed from belief systems including humanism (maximize individual control and freedom), utilitarianism and materialism (own much spend little), belief in progress (improve yourself and society), Invisible hand (provenance: the market solves), and property and self-propagation. These ideas enacted develop the society - and the firms - we experience. The outcomes are represented using chemical organization theory as

a preliminary pet model [Dittrich, Speroni, di Fenizio 2007]. The firm is modeled as a complex of coherent ideas co-evolving with other bodies of culture in society.

4 Conclusions

If firms are autopoietic systems then they are operationally closed. The set of ideas it operates remains stable: roughly speaking does it consume the ideas it produces and vice versa. The firm perpetuates its autonomous existence through individual peoples enactments of those ideas. Enaceted ideas not socially acceptable are selected against. As autopoietic systems firms are themselves autonomous and cognitive, instead of peoples mere instruments. Contrary to *communis opinio* firms are not controlled by people, not in the same way that a hammer is controlled by a carpenter. This implies that the autonomy of people is not binary but relative to that of the firm: some autonomy is relinquished and other is gained. This is consequential for the relation between firms and representants of different ideas, their stakeholders. This determines what can be expected of interventions by people into the operations of a firm.
These outcomes are relevant for a scientific audience including representatives of economics, business and management science. In addition they are relevant for people in general associated with firms, virtually everyone on Earth. Even retreating to a cabin in Alaska does not mitigate dependency. Access to this hypothesis enables a re-assessment of their relations to the firms they do business with. Just because beliefs of people in a wider cultural sphere are at the basis of firms, the presented insights can contribute to a more sustainable future.

References

1. Alchian, A.A. . Uncertainty, Evolution, and Economic Theory . The Journal of Political Economy Vol 58, No 3 pp. 211-221 . The University of Chicago Press . 1950
2. Ashby, W. R. . Principles of the self-organizing system, in Principles of Self-Organization: Transactions of the University of Illinois Symposium, H. Von Foerster and G. W. Zopf, Jr. (eds.) . Pergamon Press: London, UK . 1962 . pp. 255-278.
3. Deleuze, G. . Difference and Repetition (translation Paul Patton) . Columbia University Press New York (Eng. The Athlone Press Ltd.) . 1994 . ISBN: 0-231-08158-8
4. Dittrich, P., Speroni di Fenizio, P.: Chemical organization theory. Bull. Math. Biol. 69 (2007) 11991231
5. Luhmann, N. . Social Systems (translated by John Bednarz, Jr. with Dirk Baecker) . Stanford University Press, Stanford, CA . 1995 . ISBN 0-8047-2625-6 (pb.)
6. Maturana, H.R. and Varela, F.J. . 1972 . Autopoiesis and Cognition The Realization of the Living (orig. De Machinas y seres vivos) . 1976 . D. Reidel Publishing Company . ISBN 90-277-1016-3
7. Simon, H.A. . The Architecture of Complexity . Proceedings of the American Philosophical Society, Vol 106, No 6, pp. 467 - 482 . 1962
8. Varela, F.J., Thompson, E., Rosch, E. . The embodied mind: Cognitive science and human experience. MIT Press . 1992 . ISBN978-0262261234

Automation and occupational mobility: A data-driven network model

R. Maria del Rio-Chanona ,[1,2*] Penny Mealy,[1,4,5] Mariano Beguerisse-Díaz,[2] François Lafond,[1,3,4] and J. Doyne Farmer[1,2,6]

[1] Institute for New Economic Thinking at the Oxford Martin School, University of Oxford
rita.delriochanona@maths.ox.ac.uk
[2] Mathematical Institute, University of Oxford
[3] School of Geography and Environment, University of Oxford
[4] Bennett Institute for Public Policy, University of Cambridge
[5] Santa Fe Institute

1 Introduction

Recent studies have raised concerns about job displacement due to automation [1, 2]. However, history suggests that while some jobs are automated, new jobs are created, making it essential to understand job transitions. Here we study how automation affects employment through a model that focuses on the process of *labor reallocation*. Consider, for example, employment security faced by a statistical assistant vs. a childcare worker. Forecasts suggest that the statistical assistant is more likely to be replaced by software technology than the childcare worker [1]. However, the statistical assistant's current skills allow her to transition into occupations with low risk of automation and growing demand. In contrast, as automation displaces workers, many of them may have the skills required for childcare jobs, and may consequently threaten the job security of existing childcare workers. Thus, even though the direct replacement effect of automation is larger for statistical assistants, when we account for possible occupational transitions and labor demand reallocation, the negative impact of automation is worse for childcare workers.

2 Results

We study the propagation of an automation shock on a network-based labour market model. We first construct an occupational mobility network representing the ease with which a worker can transition between occupations. To do this we follow the work of Mealy et. al. [3], where they empirical data on occupational transitions. The Occupational Mobility Network (see Fig. 1) is weighted, directed and has self-loops. We represent the network by its adjacency matrix A with components:

$$A_{ij} = \begin{cases} r & \text{if } i = j \\ (1-r)P_{ij} & \text{if } i \neq j. \end{cases} \tag{1}$$

where the P_{ij} is $ij-$th entry of Transition matrix as defined by Mealy et. al. [3] and is the probability that a transitioning worker from occupation i moves to occupation j.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The self-loops of the network have weight r and are the probability that a worker, who is changing jobs, remains in her original occupation.

We then model the dynamics of employment ($\bar{e}_{i,t}$), unemployment ($\bar{u}_{i,t}$), and vacancies ($\bar{v}_i$) for each occupation i. In our model, workers are separated form their jobs with probability δ_u and vacancies are opened with probability δ_v. Additionally, more workers (vacancies) are separated (opened) at a rate γ_u (γ_v) depending on the target demand of each occupation $d^{\dagger}_{i,t}$. We comput the target demand of each occupation using current estimates of the automatability of the occupation [1, 2].

We assume that the skills of the workers determine their possibilities for applying to jobs and consider that skills of the worker are given by their most recent occupation. If unemployed, a worker can apply to job vacancies of her last occupation or of similar occupations. Then, we implement a search and matching mechanism with which workers take job vacancies. The dynamics of our model can be approximately described by the following equations.

$$\bar{e}_{i,t+1} = \bar{e}_{i,t} - \underbrace{\left(\delta_u \bar{e}_{i,t} + (1-\delta_u)\gamma_u \max\left\{0, \bar{d}_{i,t} - d^{\dagger}_{i,t}\right\}\right)}_{\text{separated workers}} + \underbrace{\sum_j \bar{f}_{ji,t+1}}_{\text{hired workers}}, \tag{2}$$

$$\bar{u}_{i,t+1} = \bar{u}_{i,t} + \underbrace{\left(\delta_u \bar{e}_{i,t} + (1-\delta_u)\gamma_u \max\left\{0, \bar{d}_{i,t} - d^{\dagger}_{i,t}\right\}\right)}_{\text{separated workers}} - \underbrace{\sum_j \bar{f}_{ij,t+1}}_{\text{transitioning workers}}, \tag{3}$$

$$\bar{v}_{i,t+1} = \bar{v}_{i,t} + \underbrace{\left(\delta_v \bar{e}_{i,t} + (1-\delta_v)\gamma_v \max\left\{0, d^{\dagger}_{i,t} - \bar{d}_{i,t}\right\}\right)}_{\text{opened vacancies}} - \underbrace{\sum_j \bar{f}_{ji,t+1}}_{\text{hired workers}}, \tag{4}$$

where $\bar{f}_{ij,t+1}$ is the flow of workers and $\bar{d}_{i,t} = \bar{e}_{i,t} + \bar{v}_{i,t}$ is the realized demand. Given a set of time series for the target labor demand $d^{\dagger}_{i,t}$ and a set of initial conditions, Eqs. (2 – 4) determine the expected employment, unemployment and vacancies as a function of time.

In our results we study occupation-specific unemployment and long-term unemployment (27 or more weeks) and demonstrate that our model reproduces the Beveridge Curve. Workers in highly automated occupations are more likely to be unemployed or to stay unemployed for a long period. The network structure plays an important role – workers in occupations with a similar degree of automation can have fairly different outcomes, depending on the position of the occupation in the mobility network. Automation may cause bottlenecks in the mobility network, with workers unable to find jobs for long periods. Our work highlights that retraining schemes must be directed towards workers in occupations with limited possibilities to transition to new occupations [4].

Summary. Many existing jobs are prone to automation, but since new technologies also create new jobs it is crucial to understand job transitions. Based on empirical data we construct an occupational mobility network where nodes are occupations and edges

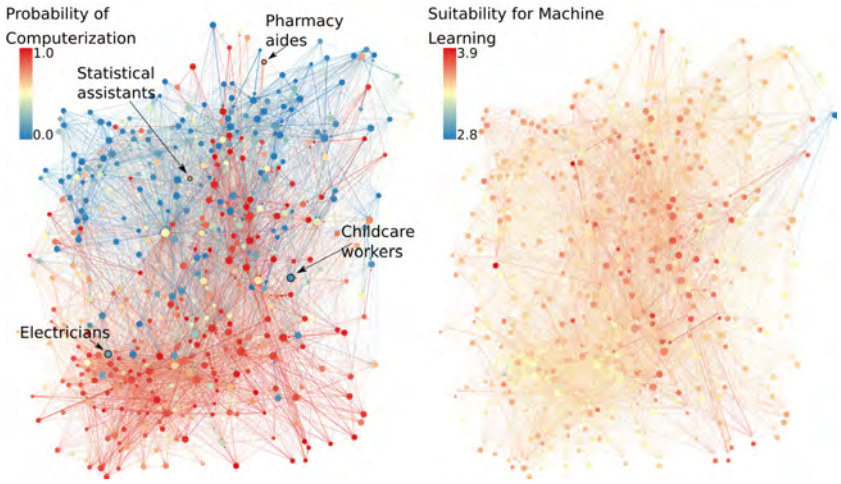


Fig. 1. Estimates of automatability in the occupational mobility network. The panels show the occupational mobility network, where nodes represent occupations and links represent possible worker transitions between occupations. The nodes are colored by: **Left)**the probability of computerization estimated by Frey and Osborne [1] and the **Right)** suitability for machine learning as estimated by Brynjolfsson et al. [2]. Red nodes have high automatability and blue nodes have low automatability. The size of the nodes indicates the number of employees in each occupation. Figure adapted from [4].

represent the likelihood of job transitions. To study the effects of automation we develop a labour market model. At the macro level our model reproduces the Beveridge curve. At the micro level we analyze occupation-specific unemployment in response to an automation-related reallocation of labour demand. The network structure plays an important role: workers in occupations with a similar automation level often face different outcomes, both in the short term and in the long term, due to the fact that some occupations offer little opportunity for transition. Our work underscores the importance of directing retraining schemes towards workers in occupations with limited transition possibilities [4].

References

1. Frey CB, Osborne MA. The future of employment: How susceptible are jobs to computerisation?. Technological forecasting and social change. 1;114:254-80 (2017).
2. Brynjolfsson E, Mitchell T, Rock D. What can machines learn, and what does it mean for occupations and the economy?. InAEA Papers and Proceedings Vol. 108, pp. 43-47 (2018).
3. Mealy P, del Rio-Chanona RM, Farmer JD. What you do at work matters: New lenses on labour. What You Do at Work Matters: New Lenses on Labour (2018). Available at SSRN: https://ssrn.com/abstract=3143064 or http://dx.doi.org/10.2139/ssrn.3143064
4. del Rio-Chanona RM, Mealy P, Beguerisse-Daz M, Lafond F, Farmer JD. Automation and occupational mobility: A data-driven network model. arXiv preprint arXiv:1906.04086. (2019).

Bow-tie Structure and Community Identification of Global Supply Chain Network

Abhijit Chakraborty[1] and Yuichi Ikeda[2]

[1] The University of Hyogo, Kobe 650-0047, Japan,
[2] Kyoyo University, Kyoto 606-8306, JAPAN,
abhiphyiitg@gmail.com / ikeda.yuichi.2w@kyoto-u.ac.jp

1 Introduction

National economies are linked by international trade and consequently economic globalization forms a giant economic complex network with strong links, i.e., interactions due to increasing trade. Especially if we view the globalized world economy with high resolution or microscopic view, we might notice that the giant economic network is a global supply chain consists of a huge number of firms. Hearnshaw *et. al.* [1] have studied the supply chain network in terms of complex network approach and have proposed nine propositions. The nine propositions are related to path length, power-law degree distribution, clustering coefficient, preferential attachment growth mechanism, truncated power-law connectivity distribution, power-law distribution of node strength, community structure with overlapping boundaries, resilience against random failure and targeted attack, core-periphery structure. They tried to explain various functions of the supply chain by the structural characteristics of supply chain network.

Here, we focus on topological properties of global supply chain network. The study on topological properties of global supply chain network is the first step to understand the globalized world economy with microscopic view. We uncover the community structure of the network using map equation method [3] and characterized them according to their location and industry classification. Furthermore, the composition of communities in terms of the bow-tie components is analyzed.

2 Results

The global supply chain data was constructed by collecting various company data from the web site of Standard & Poor's Capital IQ platform in 2018. The data include company ID, company name, country and location of company, company type, and primary industry as node information. The data also include types of business relationship between supplier and customer as link information.

As the supply chain network is directed in nature, one can define in and out degrees for the nodes. We observe probability density distributions for both nodal in and out degrees have a heavy tail nature where the tail of the distributions is characterized by a power law of the form $P(k_{in/out}) \sim k^{-\gamma_{in/out}}$ with $\gamma_{in} = 2.42$ and $\gamma_{out} = 2.11$ respectively. The high asymmetry in degree distribution can results system wide aggregate fluctuation due to idiosyncratic shocks to large firms. We observe the clustering coefficient in

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the supply chain network is a decaying function of degree having a form $\langle C(k)\rangle \sim k^{-\beta_k}$ with $\beta_k = 0.46$ indicates the presence of a hierarchical structure. Furthermore, the average degree of the neighbors of a node, $\langle k_{nn}(k)\rangle$, does not depend on k and remain more or less in constant with k, indicating the absence of nodal degree-degree correlation in global supply chain network.

We study the connected components when the network is viewed as an undirected network. The largest connected component of the network is known as the Giant weakly connected component (GWCC). The network consists of a very large GWCC with $N = 407,527$ nodes and $L = 927,316$ links. While the GWCC contains 93.16% of nodes of the network, the rest of components are very small. The bow-tie structure is uncovered from the GWCC based on the flow of goods and services (money flows in the opposite direction) along the directed links. The definitions of the different regions of the bow-tie structure are given as follows:

- The Giant strongly connected component (GSCC): The largest region where any two nodes are reachable through directed path.
- IN components: The nodes from which GSCC is reachable through directed paths.
- OUT components: The nodes that are reachable from the GSCC through directed paths.
- Tendrils (TE): The rest of the nodes in the GWCC.

The number of firms in each component is shown in Table 1. The OUT component is the largest component and it consists 41.1% of total firms. GSCC, IN and TE are approximately similar in size and comprise 16.4%, 22.3%, and 20.2% of total firms respectively. This exhibit sharp contrast with bow-tie structure observed in the Japanese production network [2].

Table 1. Bow-tie structure: Sizes of different components

Component	Number of firms	Ratio (%)	Ratio in Japanese GWCC [2] (%)
GSCC	66,798	16.4	49.7
IN	90,992	22.3	20.6
OUT	167,509	41.1	26.2
TE	82,228	20.2	3.5
Total	407,527	100	100

"Ratio" refers to the ratio of the number of firms to the total number of firms in GWCC.

We uncover the community structure of the network using map equation method [3] and characterized them according to their location and industry classification. We study the significant overexpression of different attributes such as company type, primary industry, firm's location, bow-tie components within the communities. Various interesting features can be observed from the results of attribute overexpression. The largest community comprises of private companies mainly from automotive retail based in the US. These firms belong to the OUT component in the bow-tie structure of the global supply chain network. It indicates the retail firms generally belong to the OUT component

of bow-ties structure. We construct a weighted and undirected network of countries from their overexpression in communities. A link of weight 1 is placed between two countries if they over-express simultaneously within a community. We show the over-expression network of countries in Fig. 1. Geographical dimension is clearly observed in this figure.

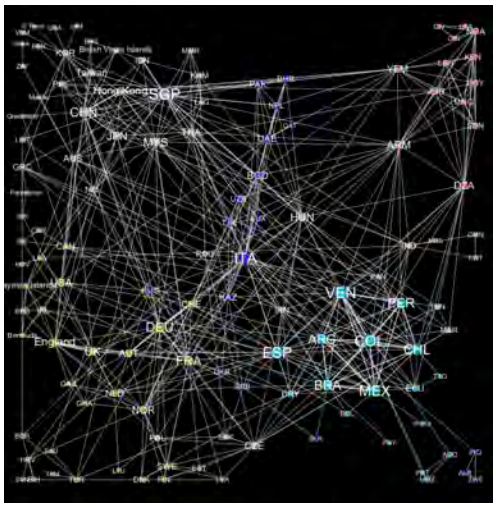


Fig. 1. Overexpression network of countries

3 Summary

We studied bow-tie structure of the giant weakly connected component in the global supply chain network. It turned out that the OUT component is the largest component. GSCC, IN and TE are approximately similar in size. We uncovered the community structure of the network. The largest community comprised of private companies mainly from automotive retail based in the US. These firms belong to the OUT component in the bow-tie structure.

References

1. Hearnshaw, E. and Wilson, M.: A complex network approach to supply chain network theory. International Journal of Operations & Production Management 33, 4, 442–469 (2013)
2. Chakraborty, A., Kichikawa Y., Iino, T., Iyetomi, H., Inoue, H., Fujiwara, Y., and Aoyama, H.: Hierarchical communities in the walnut structure of the Japanese production network. PLoS ONE 13(8): e0202739.(2018)
3. Rosvall, M. , Axelsson, D., and Bergstrom, C.T.: The map equation, Eur. Phys. J. Special Topics 178, 1323 (2009)

Evolution of users behavior in bitcoin and ethereum transaction networks

Ayana Aspembitova[1,2], Ling Feng[2], Valentin Melnikov[1], and Lock Yue Chew[1]

[1] Nanyang Technological University, Singapore
lockyue@ntu.edu.sg,
[2] Agency for Science, Technology and Research, Singapore
felney@gmail.com

1 Introduction

The invention of the blockchain and cryptocurrencies opened a great opportunity to study closer financial transactions, as this kind of information was not previously available due to its sensitivity. Cryptocurrency data open to the public from blockchain enabled researchers to construct and analyze transaction networks, user activity, money flow, etc. Several studies have contributed to the understanding of bitcoin's structure, evolution [1], [2] price formation [3], etc. There is less research done on the ethereum network, but still study of Chen [6] showed some similarities with bitcoin. The problem of user structure in the cryptocurrency market was investigated from the anomaly detection point of view. Pham and Lee [4] used unsupervised machine learning and outlier detection methods to detect suspicious nodes. However, the behavior of users in the whole cryptocurrency market is not well understood yet. In this paper we aim to investigate the user behavior composition in different periods of the cryptocurrency system - when the price was increasing, decreasing, in a stable period and the price before big fluctuations. We construct transaction networks in different periods for both bitcoin and ethereum systems, and analyze their structure and evolution. Then, we calculate various properties for all the nodes in our networks (features) and implement unsupervised machine learning method to find clusters of users with different behavior. We have been able to detect users with distinct behavior in both bitcoin and ethereum. It has been found that user behavior composition is more stable in ethereum regardless of the period, while in bitcoin the number of clusters changes significantly and different periods show different user composition.

2 Data and Methods

2.1 Data Description and Network Construction

Bitcoin transaction data was extracted from the full Bitcoin blockchain starting from the genesis block (dated 3 January 2009) up to block 560000 (dated 25 January 2019). It was then processed with BitIodine software which implements clustering of addresses into those hypothetically belonging to the same user based on two heuristics - several addresses transacting to one account are considered to belong to one user, and

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the address of a transaction which appears to be a change transaction is considered to belong to the sender of the transaction. Based on the processed data, the temporal network of interactions of bitcoin users was estimated. Since the clustering algorithm is heuristics-based, it does not guarantee that all the wallets in the network are clustered to corresponding users. Therefore, we might expect a certain fraction of non- or poorly clustered wallets, but still this algorithm results in the significant improvement of network representation of financial interaction in bitcoin. Difference between ethereum and bitcoin blockchains is that in ethereum's nodes' balances are stored directly in an account. Therefore, when obtaining data from ethereum blockchain we do not need to do the deanonymization procedure that was necessary for bitcoin.

We then construct temporal, weighted, directed networks where each link (i, j, w, t) is a transaction between two nodes (users) i and j at time t with the amount of coins w.

2.2 Methods and Feature Extraction

To detect user clusters in networks, we use k-means algorithm that comprises of the following steps [5]:

1. Initialize cluster centroids $\mu_1, \mu_2, ...\mu_k$. Optimal number of clusters k was calculated using the elbow method [8]
2. Segment data into k groups assigning each data point to the closest centroid and changing the centroid to the average of its assigned points so that the distortion function would converge:

$$J(c, \mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

where $x^{(i)}$ is the data point and $\mu_{c^{(i)}}$ is the cluster centroid

In order to use k-means as a clustering method, for each node in the network we define a set of features:

- in-degree of a node i in time interval T - total number of interactions where the node was a receiver of transaction value.
- out-degree of a node i in time interval T - total number of interactions where the node was a sender of transaction value. These two features (in and out degrees of a node) help us to get an understanding of how active a node is in sending/receiving coins.
- outgoing value - total amount of eth/btc sent in time interval T (sum of weights w of a node i when it was a sender of transaction value).
- incoming value - total amount of eth/btc received in time interval T (sum of weights w of a node i when it was a receiver of transaction value).
- total balance - net number of coins in the account balance of node i in time interval T. These three features (incoming and outgoing values and balance of a node) show the wealth of a node and its preference to accumulate or spend coins.

– size of a community that a node (user) belongs to - communities in the networks were detected using the Louvain algorithm [7]. The size of each community was calculated and assigned to nodes in that community. This additional feature was introduced in order to get better understanding of the node's position in the network. We were not able to calculate clustering coefficients and betweenness centrality for nodes in the network since the time complexity for these properties in a large network is extremely high. Therefore, we use the size of community as a feature to describe the position of a node in the network.

After calculating features for all nodes we then use Principal Component Analysis to reduce features dimension for better visualization and further analysis using the k-means clustering algorithm.

3 Results and Discussion

Figures **??** and **??** show the change of clusters in different periods of price evolution. For each period we perform user structure analysis by first, extracting features for each node and reducing their dimension by mapping them to principal components. Finally, using k-means clustering algorithm to detect user clusters as described in the Methods section.

We then do an inverse transformation from principal components back to feature space and derive the properties of users in each cluster. In the bitcoin system, in the stable period and period of price growth, there is a cluster of users with high in-degree and zero out-degree that hold around 90% of the total wealth in the network (we would speculate it might be the cluster of investors). Another cluster with distinct behavior consists of users with high out-degree and low in-degree. The third cluster in these periods have users with mixed in and out degrees and relatively low balance. In the period of price decline there is no cluster that behaves like investors, and user composition becomes more heterogeneous. In the ethereum system we see a constant number of clusters despite the period. Moreover, the behavior of users in each cluster does not change significantly in different periods as well as number of users in each cluster.

Overall, despite having similar network structure (like degree distribution and some global properties as it was shown in [1], [2] and [6]), the system evolution and behaviour of users in ethereum and bitcoin appeared to be very different. In the future, we would like to add more periods to our analysis of user composition - before big price fluctuations, after shock events etc. and also validate our speculations about the user types in each cluster. Moreover, we would like to try different methods of machine learning techniques for user clustering and compare results.

References

1. Condor D, Posfai M, Csabai I, Vattay G: Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network. PLoS ONE 9(5), (2014)
2. Aspembitova A, Feng L, Melnikov V, Chew LY: Fitness preferential attachment as a driving mechanism in bitcoin transaction network. PLoS ONE 14(8), (2019)

3. Bolt, W.: On the value of virtual currencies. SSRN Electron. J. (2016)
4. Thai T. Pham, S Lee: Anomaly Detection in Bitcoin Network Using Unsupervised Learning Methods, Unpublished report, (2016)
5. Singh A, Yadav A, and Rana A. Kmeans with three different distance metrics. International Journal of Computer Applications, 67(10):13-17, (2013)
6. Chen T, Zhu Y, Li Z, Chen J, Li X, Luo X, Lin X, Zhange X: Understanding Ethereum via Graph Analysis. IEEE Conference on Computer Communications (2018)
7. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. : Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10: P10008, (2008)
8. Kodinariya TM, Makwana PR: Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6), (2013)

Welfare effects of network structure – some results from a link formation model on monopolistic markets

Tamás Sebestyén[1] and Dóra Longauer[2]

[1] University of Pécs, Faculty of Business and Economics
Pécs, Rákóczi út 80., Hungary,
sebestyent@ktk.pte.hu
[2] University of Pécs, Faculty of Business and Economics
Pécs, Rákóczi út 80., Hungary,

1 Introduction

When modeling market mechanisms, standard economic models frequently assume that information flows on markets are frictionless, or perfect. This means that we work with models which assume that market actors are perfectly aware of other actors, their decisions (behavior) and other aggregate conditions which may be relevant for making optimal decisions. Although this seems to be a very strict assumption to work with, most economic models do not step apart from this.

A prominent example of these models is the monopolistic competition model, which was initiated by the work of Dixit and Stiglitz [7] and then developed further to become a key ingredient in different fields of applied economic modeling (see e.g. [5]). The monopolistic competition paradigm serves as the basis for the very popular New-Keynesian models (see e.g. [9]), but it constitutes the starting point for endogenous growth models ([13], [10]), models in international trade ([6], [8]) or the famous center-periphery model of Krugman [12].

In the monopolistic competition model suppliers compete with each other with prices while having some monopolistic power which in principle allows for heterogeneous prices to be determined. However, the symmetry of the actors and the assumed complete information and complete connectedness (all suppliers know all buyers and vice versa) result in homogeneous prices and a perfect awareness of actors of the average price on the market.

On the other hand, recent research on complex networks has shown that real networks show particular structures, which are far from complete in general ([4], [3]), while socio-economic systems are frequently characterized by scalefree properties. Moreover, it is also clear that the very structure of complex networks has significant effect on the performance of the whole system, especially their stability and efficiency ([11], [2], [3]).

Motivated by the previous ideas, in this paper we challenge the standard view of modeling markets by assuming incomplete connectedness among market actors. This incomplete connectedness results in a biased perception about market prices and heterogeneous prices. In addition to modeling this incomplete flow of information, we take account of the endogenous formation of supplier-buyer networks, showing that equilib-

rium networks are not necessarily complete and that under realistic circumstances, they show scalefree properties.

2 Results

In order to show the significance of incomplete information flows on market outcomes, we include a slight modification in the standard monopolistic competition model: instead of assuming that all suppliers are connected to all buyers and vice versa, it becomes the firms' endogenous decision to establish links with households depending on the possible gains coming from the increased demand and the love of variety effect, while they incur some cost of link formation/maintenance. That is, the equilibrium network itself is the result of firms' individual decisions about their network position (embededness) on the basis of a cost-benefit analysis.

The basic idea behind our model is that if buyers do not interact with all suppliers, the former have a specific information set on market prices to base their demand decisions on. Rational firms take this information into account being aware that they supply different segments of the whole market. Depending on possible additional profits, they can increase or decrease their market coverage, while considering the consequence of link formation on the behavior of buyers. While suppliers may concentrate directly on their own sub-market, if the market network is connected, other sub-markets will influence their decisions indirectly.

Using this framework we derive the optimal pricing and link formation decision of firms and analyze the resulting network structures together with price heterogeneity and the efficiency of the resulting market outcomes.

1. Incomplete network structures result in deadweight loss, an efficiency gap compared to standard market models. This gap comes from the monopolistic forces awakened by incomplete connectedness: less information on the buyers' side about competitors (coming from the lower connectedness of the former) mean that suppliers can act more like monopolists on their sub-markets. This deadweight loss is economically significant, depending on the substitutability of the products, it can reach 100% of the complete-network utility level. Moving from a Poisson to a more asymmetric degree distribution welfare may increase or decrease, depending on the alignment between the degree and productivity distribution of firms.
2. Apart from extreme situations like homogeneous products or zero cost of link formation, endogenous network formation results in incomplete network structures: it is not optimal for suppliers to cover all their markets, neither is it optimal for buyers to consider the products of all suppliers even if they prefer variety. This is important, because it calls the attention to the fact that we can not assume complete connectedness in market models as an equilibrium outcome. It can be shown that the incompleteness of the equilibrium network is a function of the cost of link formation and the substitutability of the products supplied on the market.
3. Optimal degree of suppliers depend on their productivity according to a power-law. More efficient suppliers are able and willing to attract more buyers and this relationship is not linear, meaning that a relatively small dispersion in the productivity

levels of firms is able to render the resulting equilibrium market network into a scalefree structure. This finding has an important empirical relevance as real data shows that the size distribution of firms is highly skewed and resembles the power-law distribution (see e.g. [1]).
4. While incomplete connectedness comes with a deadweight loss (inefficiency), interestingly this can be compensated by heterogeneity both in productivity levels and node degrees. We show that an increasing heterogeneity of supplier productivities results in a grouping of buyers at the more productive firms which then positively influences aggregate welfare and this way it can compensate for the welfare loss coming from the incomplete network structure.

Summary. This paper challenges the standard view of complete connectedness of market actors which is a surprisingly common assumption in most applied economic models. By simple modifications of the standard model of monopolistic competition we add endogenous link formation and analyze the resulting network structures and market outcomes. Our results show that endogenously developing market structures are generally incomplete which calls the attention to revise our knowledge about the basis and results of standard market models. Also, we show that incomplete market structures result in welfare loss due to strengthened monopolistic forces, but this can be partly compensated by heterogeneous productivity levels of the suppliers by rendering the network structure more asymmetric and providing access to better technologies for more buyers. Further research may introduce heterogeneous costs and a thorough analysis of inequality.

References

1. Axtell, R. (2001): Zipf distribution of U.S. firm sizes. Science, 293(5536) 1818-1820.
2. Bala, V. Goyal, S. (2000): A Noncoperative Model of Network Formation. Econometrica, 6(5), 1181-1229.
3. Barabsi A-L. (2016): Network Science. Cambridge University Press.
4. Barabsi A-L., Albert R. (1999): Emergence of Scaling in Random Networks. Science, 286, 509-512.
5. Brakman, S. Heijdra, B. (2004): The Monopolistic Competition Revolution in Retrospect. Cambridge University Press.
6. Dixit, A.K., Norman, V. (1980): Theory of International Trade. Cambridge University Press.
7. Dixit, A.K., Stiglitz, J.E. (1977): Monopolistic Competition and Optimum Product Diversity, The American Economic Review, 67(3), 297-308.
8. Ethier, W.J. (1982): National and International Returns to Scale in the Modern Theory of International Trade. The American Economic Review, 72(3), 389-405.
9. Gali, J. (2008): Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework. Princeton University Press.
10. Grossman, G., Helpman, E. (1991): Innovation and Growth in the Global Economy. Cambridge, MIT Press.
11. Jackson, M., Wolinsky, A. (1996): A Strategic Model of Social and Economic Networks. Journal of Economic Theory, 71(1), 44-74.
12. Krugman, P. (1991): Increasing Returns and Economic Geography. The Journal of Political Economy, 99(3), 483-499.
13. Romer, P.M. (1990): Endogenous Technological Change. Journal of Political Economy, 98(5), 71-102.

Shock propagation along the International Macronutrient Network

Marco Grassia[1], Giuseppe Mangioni[1], Stefano Schiavo[2,3], and Silvio Traverso[4]

[1] Department of Electrical, Electronics and Computer Engineering
University of Catania (Italy)
[2] School of International Studies
University of Trento (Italy)
[3] DRIC – OFCE SciencesPo (France)
[4] Department of Economics and Management
University of Florence (Italy)

The problem of global food security, which remains one of the main challenges facing humanity, is tightly connected with the structural transformation of the global food system. On the one hand, since agricultural food production globally accounts for almost 40% of land use and over 70% of water use, issues related to climate change have a direct impact on the dimension of food availability. On the other hand, the food availability, access and stability have been globally influenced by the increasing integration of international food markets. Indeed, fostered by technological factors and multilateral agreements, the amount of internationally traded food has more than doubled over the last 30 years and an increasing number of countries rely on food imports to satisfy their domestic demand. Today, about one fourth of the global food production is sold internationally.

The implications of the increasing integration of international food markets on global food security are controversial. On the one hand, it promotes a more efficient use of global natural resources and allows countries characterized by relatively unfavorable conditions for food production to fulfill their domestic food demand by specializing in the production of goods for which they have a comparative advantage. On the other hand, a highly integrated global food system may be vulnerable to systemic risks, since natural and political shocks in key countries may trigger self-propagating trade disruptions. Indeed, after the 2008 food-price spikes, calls for "food self-sufficiency" echoed in several import-dependent countries.

The paper aims at investigating the relationship between international trade and global food security from a network perspective, combining comprehensive data on bilateral trade flows in agricultural goods with network models of shock propagation. This approach allows us to analyze global food security from a systemic perspective and to shed light on the risks and trade-offs implicit in an increasingly globalized food system [1–3]. By combining information about individual country characteristics - which influence the way countries react to external shocks - with a stylized model of shock propagation, we account for direct and indirect links between countries and thus manage to study the shock-propagation mechanisms and to identify systemic fragility.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

We extend the model by [4], whereby countries respond to shocks (lower production) by imposing export restrictions, to take into account heterogeneity across countries: in particular, we postulate that importing countries are not equally affected, but rather face import restrictions that are inversely proportional to their per capital income. This assumption captures the notion that higher-income countries may be in a better bargaining position and/or can afford higher prices, and is well supported by the empirical evidence.

A second departure from the existing literature stems from the fact that, instead of focusing on a few staple food items (as it is common in the literature), we express food trade in terms of their key nutritional content (quantity of carbohydrates, lipids and proteins) and thus consider all food items that are traded internationally. This approach has two main advantages: on the one hand it allows us to describe food trade in terms of a small number of networks; on the other hand, it provides us with a clear link between international trade and food security.

References

1. Ercsey-Ravasz, Maria, and Toroczkai, Zoltan, and Lakner, Zoltan, and Baranyi, Jozsef. Complexity of the International Agro-Food Trade Network and Its Impact on Food Safety. *PLOS ONE* **7**(5): e37810, (2012), doi: 10.1371/journal.pone.0037810.
2. Sartori, Martina and Schiavo, Stefano. Connected we stand: a network perspective on trade and global food security. *Food Policy*, **57**: 114-12, (2015), doi: 10.1016/j.foodpol.2015.10.004.
3. Suweis, Samir and Carr, Joel A. and Maritan, Amos and Rinaldo, Andrea and D'Odorico, Paolo. Resilience and reactivity of global food security. *Proceedings of the National Academy of Sciences*, **112**(22): 6902–6907, (2015), doi: 10.1073/pnas.1507366112.
4. Burkholz, Rebekka, and Frank Schweitzer. International crop trade networks: The impact of shocks and cascades. arXiv preprint arXiv:1901.05872 (2019)

Community structures based on multi-attributes in International Trade Network

P. Bartesaghi[1], S. Benati[2], G.P.Clemente[3], and R.Grassi[1]

[1] Universit Milano-Bicocca, paolo.bartesaghi@unimib.it, rosanna.grassi@unimib.it
[2] Universit degli Studi di Trento stefano.benati@unitn.it
[3] Universit Cattolica del Sacro Cuore, Milano, gianpaolo.clemente@unicatt.it

1 Introduction

Community detection is a widely discussed topic in network theory (see, e.g., ([3])). The analysis of the mesoscale structure of a real network throws light on its inner structure. This plays an even more significant role when applied to International Trade Network (ITN), in view of its multiple implications (see, for instance, [1]). In this framework, this work aims at clustering countries according to similarities in their roles in the global market, rather than using only the preferential channels of exchange between them. Our contribute to the literature is to provide a new methodology of community detection that aims at grouping countries in the ITN on the basis of more than one attribute of node similarity. Network attributes mean more specifically node properties, represented through, for instance, centrality measures or interconnection. To this end, we assess the role of each country in the ITN by means of a set of network topological indicators allowing to sort the relevance of nodes in the network under different configurations. We sum up this initial multi-criteria assessment, by defining a proper measure of similarity/dissimilarity between countries using their ranking positions. Next, we cluster data in order to find groups of nations that have common features in terms of those rankings. In this way, we are able to group countries characterized by a similar relevance in the network. The main advantage is the fact that we do not focus on a specific network indicator, but we identify set of countries that have a similar behaviour in the network from different and wider perspectives. Additionally, we provide a new heuristic algorithm in order to find clusters, based on the clique partition model, first introduced in [4],[5] and [2]. The proposed heuristic overcomes problems of existing methodologies, because not only units can be inserted in a cluster, but clusters themselves can be merged with other clusters.
An empirical application to ITN in the year 2014 is provided. The optimal solution shows three big clusters, more or less equivalent in terms of number of members but very different in terms of intra-cluster density. This has been easily interpreted, since the rate of exchanges between top countries is far more intense than for poor ones. Being a rather rough partition, we iterated the same methodology to each cluster and so on. This allows to build a dendrogram tree stemming at each step by applying iteratively the same methodology. Main results show how we are able to provide different clusters where countries with a similar relevance in the network are grouped.

2 Main proposal

In this section, we describe a new and more general approach that aims at grouping countries in the ITN on the basis of more than one attribute of node similarity. Network indicators are numerical values strictly related to the topology of the network and they are often useful in quantifying

its properties. To this aim, we consider a weighted and directed graph $D = (V,E)$ where V and E are respectively the set of n vertices and m arcs. In particular, each country in the network is represented by a vertex, while trade input and output flows are described by arcs. In our analysis, we take advantage of different indicators of vertex importance and interconnection (namely, in and out strength, in and out clustering coefficient, hub and authority, in and out Laplacian centrality). Each measure has peculiarities and characteristics that highlight various aspects of the exchange relations between countries, capturing in an exhaustive way their complexity. This heterogeneity requires an approach that cannot be simply based on the direct comparison among extremely different measures. To this reason, we focus the analysis on the rankings instead on the absolute values. Indeed, each indicator induces a ranking which represents the structural importance of a single node in the network. The comparison is developed by computing a distance function between rankings. In particular in this work we refer to the Minkowski distance, also known as L_p-norm distance. Let us order the scores of nodes obtained for each centrality measure k and let r_i^k be the position of the node i with respect to k. The Minkowski distance $d(\mathbf{r}_i, \mathbf{r}_j)$ is

$$d(\mathbf{r}_i, \mathbf{r}_j) = ||\mathbf{r}_i - \mathbf{r}_j||_p = \left(\sum_{k=1}^{K} \left| r_i^k - r_j^k \right|^p \right)^{1/p} \tag{1}$$

being $\mathbf{r}_i$ the rankings vector of node i, K the number of considered centrality measures and p any real value such that $p \geq 1$.

We use this distance to construct a complete network K_n having the same node set and weighted adjacency matrix Ω, whose entries are defined as:

$$\omega_{ij} = \begin{cases} \frac{1}{1+d(\mathbf{r}_i,\mathbf{r}_j)} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}. \tag{2}$$

These weights range in $[0;1]$ and turn out to be effective in describing the similarities between countries. Indeed, more two countries have a similar behaviour, more the distance is small, and then the weight is high. At this point, we define the integer linear programming formulation of the Clique Partition as:

$$\max \sum_{i \neq j} g_{ij} x_{ij} \tag{3}$$

subject to

$$\begin{cases} -x_{ij} + x_{ik} + x_{jk} \leq 1, & \forall i < j < k, \;\; i,j,k \in V \\ -x_{ik} + x_{jk} + x_{ij} \leq 1, & \forall i < j < k, \;\; i,j,k \in V \\ -x_{jk} + x_{ij} + x_{ik} \leq 1, & \forall i < j < k, \;\; i,j,k \in V \\ x_{ij} \in \{0,1\}, & i < j, \;\; i,j \in V \end{cases}.$$

Notice that x_{ij} is equal to 1 if two nodes are in the same cluster and 0 otherwise. Each gain/cost g_{ij} is defined as the difference between the actual and the hypothetical similarity: $g_{ij} = \omega_{ij} - 2\frac{\omega_i \omega_j}{\omega}$, with $\omega = \sum_{ij} \omega_{ij}$ the total network similarities and $\omega_i = \sum_j w_{ij}$ the sum of similarities allocated to unit i.

A specific tool developed for our project is a new heuristic algorithm in order to find clusters, based on the Clique Partition model. Indeed, we experimented very long computational times when we tried to solve it through Integer Linear Programming. Therefore, we implemented a heuristic procedure based on shrinking the vertices of the graph. We found that the algorithm calculates quickly good quality solution. However, it can be the case that the selected partition is suboptimal. Therefore, we also integrated the algorithm with a version of the Neighborhood Search procedure.

To conclude, we can summarize our proposal in the following way:

1. Given a graph D with n nodes and m edges, we select a set of K centrality measures and we compute the ranking r_i^k for the centrality measure $k = 1, 2, ..., K$;
2. For each couple of countries, we calculate the Minkovsky distance $d(\mathbf{r}_i, \mathbf{r}_j)$ via formula (1);
3. We construct a new complete network K_n having the same node set and weighted adjacency matrix Ω, whose entries are defined in formula (2);
4. We calculate the gain/cost g_{ij} based on the difference between the actual and the hypothetical similarity
5. We provide a new heuristic algorithm and we solve the Clique Partition model (3) whose input are g_{ij} 's.

3 Main results

For the sake of brevity, we report here only main results obtained solving problem (3) by using ITN data in 2014. As shown in Figure 1, the initial breakdown in communities gives a general feeling of the relevance of different macro-regions in the whole trade network. Indeed the top cluster, characterized by 69 countries at step 1, includes all the most developed European countries, largest economies in Asia and Middle East, several countries in South America, Canada, Mexico, USA, Australia and New Zealand. Except for some small countries, this community includes all the advanced economies identified in the World Economic Outlook (WEO) by the International Monetary Fund (IMF) and the emerging economies identified by IMF and by other analysts.
The following steps produce a more granular division of countries. At the end of the procedure, we obtain that the most central group is composed by China, Germany, Japan and United States. Higher volumes of trades are indeed moved by this country and at the same time, they also show highest levels of interconnections.
In the second group, we have countries which either are positioned at a slightly lower level (as GBR, FRA, ITA and NLD) or are outstanding for one specific indicator, but, on average, shows a less relevant role in the network. For instance, Canada has the second position in terms of hubs centrality, but shows an average ranking around 14, because of a lower clustering.

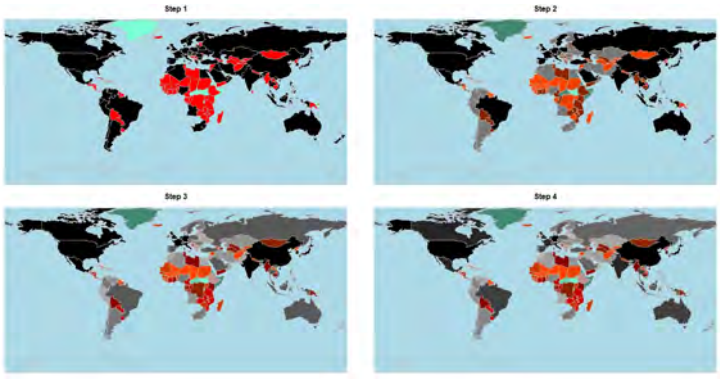


Fig. 1. Structure of communities at different steps. Darker colours are associated to communities with an higher average ranking. The number of communities is respectively equal to 3, 8, 16, 22.

References

1. Matteo Barigozzi, Giorgio Fagiolo, and Giuseppe Mangioni. Identifying the community structure of the international-trade multi-network. *Physica A: statistical mechanics and its applications*, 390(11):2051–2066, 2011.
2. S.G. de Amorim, J.-P. Barthélemy, and C.C. Ribeiro. Clustering and clique partitioning: Simulated annealing and tabu search approaches. *Journal of Classification*, 9(1):17–41, 1992.
3. Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
4. M. Grötschel and Y. Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45(1-3):59–96, 1989.
5. M. Grötschel and Y. Wakabayashi. Facets of the clique partitioning polytope. *Mathematical Programming*, 47(1-3):367–387, 1990.

Similarity and systemic risk in the network of mutual fund holdings

Danilo Delpini[1,2], Stefano Battiston[3], Guido Caldarelli[2,4], and Massimo Riccaboni[2,5]

[1] University of Sassari, Dept. of Economics and Business, Sassari, Italy,
ddelpini@uniss.it,
[2] IMT School for Advanced Studies, Lucca, Italy,
[3] University of Zurich, Dept. of Banking and Finance, Zurich, Switzerland,
[4] ISC-CNR Uos Sapienza, Rome, Italy,
[5] KU Leuven, Dept. of Managerial Economics, Strategy and Innovation, Leuven, Belgium

1 Introduction

Diversification is a well-understood strategy to reduce unsystematic portfolio risk but its systemic implications are less clear. During periods of high volatility more stocks are required to diversify effectively [1] but more generally it is not obvious how to achieve optimal diversification. During crises many investors exit risky positions and enter positions into what are considered safe investments [2, 3] and it is important to understand what happens when investors allocate similar portfolios. We move from the hypothesis that, while diversification reduces risk at the portfolio level, the *differentiation* of the investments across portfolios reduces risk at the systemic level.

2 Results

We consider US equity mutual funds as a case study of major interest [4]. We parse data from the CRSP Mutual Funds Bias-Free Database and aggregate information on portfolio composition to recover quarterly snapshots of the network of portfolio holdings. This is represented by a weighted bipartite graph, one vertex class corresponding to funds' portfolios and the other to their assets. The edge weight $W_{i\alpha}$ stands for the total value of security α in portfolio i, the total net assets (TNA) is given by $S_i = \sum_\alpha W_{i\alpha}$ and the network value by $S_{\text{tot}} = \sum_{i,\alpha} W_{i\alpha}$. We measure portfolio diversification in terms of the *inverse Herfindahl–Hirschman index* $h_i = [\sum_\alpha w_{i\alpha}^2]^{-1}$, with $w_{i\alpha} = W_{i\alpha}/S_i$, which estimates the number of leading assets. Portfolio overlap is measured by the *cosine similarity* $s_{ij} = \mathbf{w}_i \cdot \mathbf{w}_j / \|\mathbf{w}_i\| \|\mathbf{w}_j\|$, that considers what assets two portfolios have in common as well as their proportions.

We find that the number of different assets in the network has increased steeply over years and that portfolio diversification has slightly increased ($\approx 20\%$) across the Global Financial Crisis (GFC). The probability density function (PDF) of portfolio diversification does not provide evidence of any characteristic scale: most funds invest in few assets, some have thousands and a spectrum of intermediate behaviors exists. The network has very popular assets which are likely to be found in virtually any portfolio (hubs). Heterogeneity characterizes the portfolio similarities: the PDF extends over

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

several orders of magnitude and a probability exists for very similar, nearly identical portfolios. This suggests the existence of strong correlations between the investments of different funds.

We may suppose that large similarities are due to finite size effects (limited number of assets in the investment universe) or the presence of hubs. We exploit two null models of random investments to rule out such possibilities. In the random holdings model (RH) original holdings are re-assigned choosing assets uniformly at random. In the Degree-Preserving Random model (DPR) original holdings are shuffled by double-edge swap. Both models preserve the portfolios' TNAs and the fund degree sequence. The original values of portfolio diversification are preserved as well because the edge weights are unchanged. The degree sequence of the assets and the hubs are preserved only by model DPR. Comparison of the similarity probability distributions shows that unconstrained randomization suppresses portfolio overlap and, more interestingly, that very similar portfolios are not observed for model DPR either, despite hubs. We conclude that the similarity distribution, and especially its tail, are not compatible with random scenarios, not even when controlling for the effects of very popular assets. This is indirect evidence of non trivial correlations between investments.

To measure systemic riskiness, we introduce a basic model of propagation for exogenous shocks to asset prices, inspired by the documented flow-performance relationship for funds [5]. After negative portfolio returns individual investors ask for redemption of their fund shares and asset managers liquidate a fraction of the assets to repay leaving investors. The increase of offer has a negative market impact on asset prices, which depends on asset liquidity characteristics. We model the impact as a linear function $\delta_\alpha(t+1) = \Delta V_\alpha(t)/\lambda_\alpha$, where $\Delta V_\alpha(t)$ is the total amount of α liquidated by the end of trading period t and this results in a relative variation of the asset's price $\delta_\alpha(t+1) \leq 0$. The parameters λ_α are called the *market depth* of the assets and depend on the number of outstanding shares. Assuming the total value of the asset as a proxy of liquidity, we also take $\lambda_\alpha \propto V_\alpha = \sum_i W_{i\alpha}$. Due to the market impact, portfolios undergo a new round of losses and, because of the overlap of investments, some portfolios that were not hit at first will be during the following periods in a cascade of negative portfolio returns and subsequent asset liquidations. The model features *contagion effects* due to common exposures as well as *heterogeneity* in asset liquidity characteristics. We perform simulations for a shock of -30% to the 10 most common assets. The dynamics is iterated over a number T of trading periods and the *systemic damage* is computed as the relative total loss of value of the portfolios $D(t) = S_{\text{tot}}(t)/S_{\text{tot}}(t-1) - 1$.

To validate the hypothesis that portfolio overlap is connected to systemic riskiness, we first look at the corresponding time series for the real network (panels A, B of Fig. 1) and actually we find they are strongly correlated. We then simulate the shock propagation and compare the values of $D(t)$ in the three cases (Fig. 1 C). Model RH has no strongly connected assets and it has weak overlap between portfolios; correspondingly it provides the most robust configuration. Most noticeably, model DPR is more fragile than RH but still it is more robust than the real network. The latter is the most risky and we see that a large shock can propagate quite rapidly. By comparison we conclude that the observed riskiness is not explained by strongly connected assets and, more generally, it depends on portfolio overlap. Since the random benchmarks also preserve the

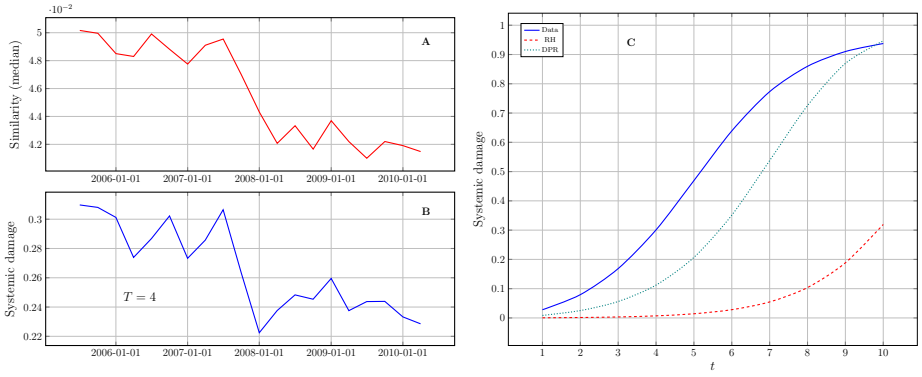


Fig. 1. Portfolio overlap (**A**) and systemic damage (**B**) are strongly correlated for the real network ($\rho = 0.91$). Comparison of the systemic damage for the real portfolios and the random benchmarks (**C**) shows that systemic riskiness is explained only partially by the effects of hubs.

original values of portfolio diversification, we also conclude that portfolio similarity is responsible for a systemic risk component that is independent of portfolio diversification and hubs.

Summary. The network of US mutual funds is a heterogenous system and during GFC portfolios have become more diversified and less similar. We find that observed similarities are more likely than expected by chance and finite-size effects. Stress tests show that the real network is risky with respect to random scenarios to parity of portfolio diversification and that riskiness can be justified by hubs only partially. More generally, a systemic risk component originates from portfolio overlap independently of popular assets and diversification. We exploit deliberately simple random models that serve the purpose of performing a comparison of the observed similarities and measuring the effects of the topology and portfolio overlap on systemic fragility. Future work will be devoted to study realistic scenarios that take into account the actual availability of shares on the market as well as the the roles of the funds' and companies' sizes in shaping investment strategies and determining portfolio similarity.

References

1. Hu JL, Chang TP, Chou RY.: Market conditions and the effect of diversification on mutual fund performance: should funds be more concentrative under crisis? Journal of Productivity Analysis 41(1), 141–151 (2014).
2. Agnew J, Balduzzi P, Sundn A.: Portfolio Choice and Trading in a Large 401(k) Plan. American Economic Review 93(1), 193–215 (2003).
3. Hurd MD, Rohwedder S.: Effects of the Financial Crisis and Great Recession on American Households. Working paper no. 16407, National Bureau of Economic Research (2010).
4. Delpini, D., Battiston S., Caldarelli G., Riccaboni, M.: Systemic risk from investment similarities. PLoS ONE 14, e0217141 (2019).
5. Fricke C., Fricke D.: Vulnerable asset management? The case of mutual funds. Bundesbank Discussion Paper No. 32/2017.

Network structure of traditional craft industry in Kyoto

Daisuke Sato[1]*, Yuichi Ikeda[1], Shuichi Kawai[1], and Maxmilian Schich[2]

[1] Kyoto University, Kyoto 606-8303, Japan,
[2] The University of Texas at Dallas, TX 75080, USA

1 Introduction

In a highly developed industrialized society, mass production using machines is at the center of economic activity. This mass production arguably stands in contrast to production methods such as the flexible division of labor [1] that characterizes the collaboration of companies and the manufacturing system of traditional craft industries in Japan. As early as 1980, such a traditional flexible division of labor has attracted attention as a substitute model to replace American-style mass production in regional economies, as for example in and around Kyoto. However, in recent years, due to changes in consumer demand, accompanied by generational transformations, the traditional craft industry in Kyoto witnesses substantially diminished sales. It seems, the hitherto effective flexible division of labor, the production equipment, the know-how, and the conventional procedures of skilled labor in practitioners of traditional crafts now increasingly become unable to cope with market fluctuations and technological changes caused by economic globalization. In this paper, we address this issue by working towards establishing a method to clarify the structure of supply chain networks between individual companies within the current traditional craft industry in the Kyoto region. Our goal is to eventually understand the underlying structures and dynamics, to eventually nurture sustainability and ensure the survival of the traditional cultural industry within the larger existing and changing market environment. As a result of our research, we expect a valuable contribution towards raising production efficiency while preserving desired qualities in traditional procedures and products.

2 Network Analysis

The data analyzed in this paper is a subset of 5,943,072 supply chain transaction relationships between 1,668,567 individual Japanese companies (including the company name, location, industry, number of employees, etc.), as investigated by Tokyo Shoko Research, Ltd in 2016. From this data, we construct a supply chain network, where each individual company is a node, and links connect pairs of companies with at least one transaction relationship, for the purpose of our present analysis resulting in a directed graph. Because of a limitation of the dataset, links have no weights. Choosing companies in Kyoto and companies that have business relationships with companies in Kyoto from the full dataset, our analysis in this paper focuses on 80,508 nodes and 153,066

*sato.daisuke.23a@st.kyoto-u.ac.jp

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

links. Given this subset of data, the field of network science offers a toolset to identify specific industrial communities and to analyze relevant topological structures of the supply chain network[2]. Specifically, to identify industrial communities, we use the map equation method, where a set of nodes is regarded as belonging to a community based on a random walker staying within the respective set of nodes with high probability [3]. Analyzing the network on the system level, we quantify the average shortest path length and assortativity. Previous research proposed the hypothesis which is the shorter the average shortest path length is, the more efficient a given the supply chain appears[2]. In terms of assortativity we quantify the degree correlation. Analyzing the network on the node level, i.e. focusing on individual companies, we quantify the IN-, OUT-, and undirected degree centralities as well as the betweenness centrality. As a first approximation, we assume nodes with high IN-degrees to play the role of integrators that design and/or process items of transaction in acts of combination. Nodes with high out-degrees, again in first approximation, are considered to distribute limited resources to consumers. Node with high betweenness centralities will perhaps play a central role in regarding the speed of supply chain flow.

3 Results

First, we identify communities in the traditional craft industry in Kyoto using a map equation. As a result, the 80,508 companies related to Kyoto is divided into 1313 communities. Each community has a hierarchical structure, which consists of subcommunities in a lower hierarchical layer. From the hierarchical structure, we chose the three communities as representative of the larger Kyoto traditional craft industry, covering the Nishijin silk fabrics industry, the Kyoyuzen dyeing industry, and the Kyoto-ningyo (i.e. Kyoto doll) industry. Fig.1 shows the embedding of all three chosen craft industries within the Kyoto supply chain network as a whole. Sub-community *3-1-1* is identified as the Nishijin silk fabrics Industry; similarly, the sub-community *3-1-30* corresponds to Kyoto Yuzen and sub-community *3-7* to Kyoto-ningyo. Here, *a-b-c* means that community *a* includes sub-community *b* and sub-community *b* include further sub-community *c*. Next, we compared the three traditional craft industries with subdivisions of modern industries, including the electronics industry and the industry of civil engineering. Table 1 indicates the topological features and centralities of the traditional and modern industrial communities. The Nishijin silk fabrics industry, it turns out, has a longer average shortest path length than other industries. This perhaps indicates that the Nishijin silk fabrics industry has a less inefficient supply chain structure. In addition, we find that there are many wholesalers with high centralities in the Nishijin silk fabrics network. As pointed out in previous qualitative research[4], the structure of wholesalers has a strong influence in the Nishijin silk fabrics industry. Common wisdom that is subject to further analysis has it that the efficiency is higher if manufacturers sell their products directly. Looking at the profits of companies with high betweenness centrality, many companies are in the red. This affects the sustainability of the industrial community, as the bankruptcy of a company with high betweenness centrality potentially has a significant impact on the entire network, pointing to a critical situation in case of the Nishijin silk fabrics Industry. The Kyoto-ningyo doll industry has the same characteristics as the

Nishijin silk fabrics industry, with a long average shortest path length, and high centrality of wholesalers. We, therefore, conclude that the doll industry likely has similar issues as the Nishijin silk fabrics industry. In contrast, manufacturing companies are more highly central in the Kyoyuzen dyeing industry, while still being subject to a long average shortest path length. We, therefore, assume that this last case ranges in-between the other two traditional and the two modern cases in terms of supply chain efficiency.

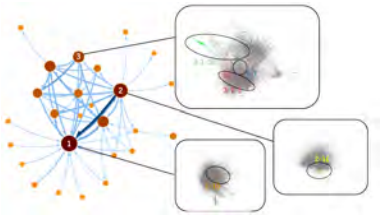


Fig. 1. Community structure of Kyoto supply chain network (Community 1:Electronics Industry, Community 2: Civil Engineering, Community 3: Traditional Industry)

	Kyoto craft traditional industries			modern indutries	
	Nishijin silk fabrics	Kyoyuzen dyeing	Kyoto doll	Electronics	Civil engineering
Community	(3-1-1)	(3-1-30)	(3-7)	(1-10)	(2-10)
Node	121	31	852	173	65
Edge	251	36	1243	259	83
Network level					
Assortativity	-0.153 (-0.051)	-0.615 (-0.127)	-0.279 (-0.113)	-0.392 (-0.234)	-0.478 (-0.220)
Average path length	1.091 (1.519)	0.874 (0.758)	0.792 (1.960)	0.348 (0.452)	0.366 (0.314)
Clustering coefficient	0.032 (0.053)	0.0 (0.057)	0.097 (0.099)	0.136 (0.059)	0.089 (0.037)
Node level					
Total degree centrality	W: 86% M: 14% D: 0%	W: 18% M: 71% D: 11%	W: 50% M: 45% D: 5%	W: 33% M: 67% D: 0%	W: 30% M: 20% D: 40%
Betweenness centrality	W: 68% M: 27% D: 5%	W: 22% M: 67% D:11%	W: 45% M: 50% D: 5%	W: 44% M: 56% D: 0%	W: 30% M: 30% D: 40%

(W:wholesalers, M:manufacturers,D:distributers)

Table 1. Comparrison of traditional industries with modern industries. Path length and clustering coefficient calculated for network with degree-preserving randomization in parentheses

4 Summary

We characterized features and issues regarding the supply chain networks of three communities of traditional craft industries in Kyoto, further compared with two modern instances. We found that the traditional craft industries in Kyoto have a longer average shortest path length than that of modern industries in their supply chain networks. We also found wholesale companies more central in two of the three traditional cases, which confirms previous qualitative research. In ongoing research, we will aim to determine how these features are related to the inefficiencies and issues underlying the known reduction of sales within the traditional industries.

References

1. Piore, M., Sabel, C.: The Second Industrial Divide Possibilitiesfor Prosperity, New Yorkz Basic Books (1984)
2. Hearnshaw, E., Wilson, M.: A complex network approach to supply chain network theory, International Journal of Operations & Production Management Vol. 33 No. 4, 442–469 (2013)
3. Rosvall, M., Axelsson, D. and Bergstrom, C.: The map equation, Eur. Phys. J. Special Topics 178, 13–23 (2009)
4. Kingo, K.: Present Conditions and Structures of the Nishijin Textile industry, Kansei Kougaku (the journal of the Japan Society of Kansei Engineering) Vol. 5 No. 9, 299–304 (1999) (in Japanese)

Economic complexity of prefectures in Japan

Abhijit Chakraborty, Hiroyasu Inoue, and Yoshi Fujiwara

Graduate School of Simulation Studies, The University of Hyogo, Kobe 650-0047, Japan
abhiphyiitg@gmail.com

1 Introduction

Every nation takes priority for inclusive economic growth and development of all regions. However, we observe economic activities are clustered in space, which results in a disparity in income per capita among different regions. A complexity based method was proposed by C. A. Hidalgo and R. Hausmann [1] for explanation of large gaps in income per capita across countries. Subsequently, A. Tacchella *et. al.* have introduced the fitness-complexity algorithm [2] based on the conceptual framework of C. A. Hidalgo and R. Hausmann to calculate the intangible properties like the fitness of countries and the complexity of export products. Although there is an extensive study on countries' economic complexity using international export data, economic complexity at regional level [3] is relatively less studied. Here, we study the industrial sector complexity of prefectures in Japan based on the basic financial information on more than a million firms. In this study, we aim to explain the economic performance of prefectures with the quantitative measure of complexity.

Our data is based on a survey conducted by Tokyo Shoko Research (TSR), one of the leading credit research agencies in Tokyo, and was provided to us through the Research Institute of Economy, Trade and Industry (RIETI). We use "TSR Kigyo Jouhou" (firm information), which contains basic financial information on more than a million firms. The data set was collected in July 2016. We only considered "active" firms that have employee and current year sales information. It contains $N = 1,033,518$ firms. We aggregate the data as a bipartite network of prefectures ($P = 47$) and industrial sectors ($S = 97$). We have used Japan Standard Industrial Classification, November 2007, Revision 12. The bipartite network is represented by the binary matrix M_{ps} , where $M_{ps} = 1$, if the prefecture p has a significant number of firms of industrial sector s, and 0 otherwise. A prefecture p is said to have a significant number of firms from industrial sector s if its Revealed Comparative Advantage (RCA) is greater than unity. The RCA is defined as [3]

$$RCA_{ps} = \frac{\frac{n_{ps}}{\sum_s n_{ps}}}{\frac{\sum_p n_{ps}}{\sum_{p,s} n_{ps}}}$$

Where n_{ps} is the number of firms in prefecture p from industrial sector s.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

The fitness-complexity algorithm is one of the quantitative methods to calculate the fitness of countries and the complexity of products. Here we use the method to study Japanese industrial sector and prefectures relationships.

Mathematically the model can be described by the following self-consistent iterative coupled equations with fitness F_p of prefectures and complexity Q_s of industrial sectors:

At any arbitrary iteration step n

$$\tilde{F}_p^{(n)} = \sum_s M_{ps} Q_s^{(n-1)}$$

$$\tilde{Q}_s^{(n)} = \frac{1}{\sum_p M_{ps} \frac{1}{F_p^{(n-1)}}}$$

with normalization in each step: $F_p^{(n)} = \frac{\tilde{F}_p^{(n)}}{<\tilde{F}_p^{(n)}>}$; $Q_s^{(n)} = \frac{\tilde{Q}_s^{(n)}}{<\tilde{Q}_s^{(n)}>}$

The initial conditions are $\tilde{Q}_s^{(0)} = \tilde{F}_p^{(0)} = 1$ for all p and s.

The convergence properties of the algorithm depends on the structure of M_{ps} [4]. We have investigated the triangular structure of the binary matrix M_{ps} by ordering the rows and columns according to their fitness complexity rank. The structure of the ordered M_{ps} shows the diagonal line does not pass through the external area, which ensures that the fitness values of the prefecture and complexity values of the industrial sectors will converge to the non zero fixed values with iterations [4]. Fig. 1 (left) shows the evolution of the prefectures' fitness with iterations. It shows that the distribution of F_c gradually broadens and finally each fitness converge to the fixed point values.

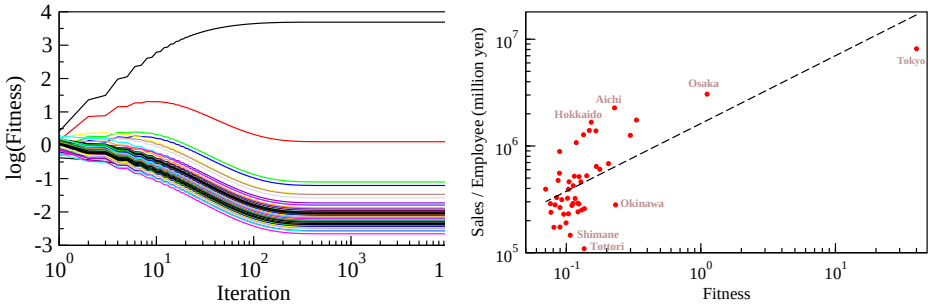


Fig. 1. The evolution of prefectures' fitness towards fixed points is plotted with iteration (left). The variations of sales per employee is plotted with fitness for the prefectures (right). The dotted line represents expected level of sales per employee, which is the best power law fit to the data with an exponent 0.63.

We show the relationship between sales per employee and fitness values for the prefectures in Fig. 1 (right). It shows that there is a strong positive correlation between the

two quantities. The Pearson's product-moment correlation between sales per employee and fitness values is found to be 0.885 with p-value $< 2.2 \times 10^{-16}$. The deviations of real sales per employee data from the expected values are informative and it gives an indication of the economic performances of the prefectures. The prefectures appearing below the expected values of sales per employee, have the potential to grow faster in the future. Furthermore, we have also observed high correlation with gross prefectural product per capita. To check the robustness of our results, we have compared it with the results obtained using Hidalgo and Hausmann method [1].

Summary. We have studied the economic complexity at prefecture level in Japan. The computed economic complexity for the prefectures shows high correlation with macro-economic indicators, such as sales per employee and gross prefectural product per capita. Further studies in this direction can predict the macro-economic indicators for a prefecture.

References

1. Hidalgo, C. A., Hausmann R.: The building blocks of economic complexity. Proceedings of the national academy of sciences 106, 10570-10575 (2009). doi:10.1073/pnas.0900943106
2. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., Pietronero, L.: A new metrics for countries' fitness and products' complexity. Scientific reports, 2, 723 (2012). doi:10.1038/srep00723
3. Gao, J., Zhou, T.: Quantifying China's regional economic complexity. Physica A 492, 1591-1603 (2018). doi:10.1016/j.physa.2017.11.084
4. Pugliese, E., Zaccaria, A., Pietronero, L.: On the convergence of the Fitness-Complexity Algorithm. The European Physical Journal Special Topics, 225, 1893-1911 (2016). doi:10.1140/epjst/e2015-50118-1

Nonparametric correlation sign prediction from high-dimensional asset price correlation matrices

Christian Bongiorno and Damien Challet

Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes
CentraleSupélec, Université Paris Saclay
3 rue Joliot-Curie, 91192, Gif-sur-Yvette, France,
christian.bongiorno@centralesupelec.fr

1 Introduction

Correlations between stock returns play a central role in finance, for example, in portfolio optimization. Buiding optimal mean-variance portfolios [6] requires very good estimations of trends, covariances, and correlations. It is has been long recognized that portfolio optimization without proper filtering is akin to error maximization [8]. The main reason is that a precise estimation of a full unfiltered correlation matrix between N assets requires $T \gg N$ points per asset. Regrettably, the non-stationary nature of financial markets imposes T to be as small as reasonably possible but in any case proportional to N. The impossibility to approximate the $T \to \infty$ limit while keeping N constant is known as the 'curse of dimensionality' as correlation estimators remain noisy even in the N and $T \to \infty$ limit at fixed ratio $q = T/N$. Ad-hoc filtering techniques include linear shrinkage [5], block-diagonal ansatz for the correlation matrix [7] and random matrix theory-based eigenvalue clipping [4, 9]. The latter works reasonably well for $T > N$. More recently, the Rotational Invariant Estimator (RIE) was shown to be optimal in the large N and T limit at constant ratio $q = T/N > 1$ [1], however, RMT and RIE assume stationary Gaussian returns and $N < T$.

Here, we focus on non-stationary correlation structures of possibly non-Gaussian returns when $N > T$ and aim to predict the sign of asset correlations.
Our approach is related to Heider balance theory [3], which aims at explaining the attitude changes of interacting individuals. In the modeling framework of this theory, only two possible interactions between two individuals are possible: the latter can be *friends* or *enemies*. The general observation in social science that 'the enemy of my friend is my enemy' becomes particularly relevant when extended to triadic relationships: for example, triads where a is a friend of b and c but c is an enemy of b tend to be unstable. As a consequence, one interaction type is likely to change and lead to a stable triad: a could become an enemy of b, or c could become a friend of b. In a similar way, a triad composed of three individuals that are enemies of each other is considered unstable as two individuals could join their forces against the third one. In summary, this theory identifies four possible triads, two stable ones and two unstable ones, and adds the intuition is that unstable triads, due to social stress, tend to evolve into stable ones.
Recently, in Ref [2], the authors proposed to measure the global social balance with a Hamiltonian whose minimal energy level coincides with the maximal stability and studied the possible paths that drive the system towards minimal energy levels, i.e., to

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

the maximally stable triad states. Inspired by their work, we defined a new metric Δ_{ij} which aims to to identify the correlation pairs that tend to revert their sign. Such a metric provides a quantitative measure of the social stress that a correlation sign induces on the whole system according to Heider balance theory.

2 Results

We validate our approach on daily adjusted close-to-close returns of equities from US and Hong Kong stock markets, for a period spanning from 2000 to 2018. According to our hypothesis, the correlations that are involved in a large number of unstable triads are the most susceptible to revert their signs (Fig. 1(a)). To assess the overall ability of Δ_{ij} to predict the sign stability, we used the Receiver Operating Characteristic (ROC) curve. We compare the ROC curves associated with Δ_{ij} as discrimination variable, and the other associated with the absolute value of the correlation $|\Phi_{ij}|$ (Fig. 1(b)). As a summary of the performance of a discrimination variable, we use the Area Under the Curve (AUC). We computed the performance both predictors for a wide range of calibration and test window lengths chosen in order to include partial-rank and full-rank correlation matrices. Although the difference between the methods is not constant over time (Fig. 1(c)), our approach outperforms $|\Phi|$ whenever $N > T$ (Fig. 1(d)).

Summary. In the high-dimensional regime, correlation matrices become pathologically noisy, and their coefficients cannot predict when their sign is likely to change. Using fewer bits of information per price return but accounting for more complex relationships between correlations makes it possible to predict the sign change of correlation coefficients deep in the high-dimensional region, even when there are ten times fewer data points than assets. More precisely, triadic relationships suggest a stability measure of each sign of correlations which was shown to outperform the absolute value of correlations. In short, higher-order nonparametric structures lift the degeneracy of the high-dimensional correlation matrices.

References

1. Bun, J., Allez, R., Bouchaud, J.P., Potters, M.: Rotational invariant estimator for general noisy matrices. IEEE Transactions on Information Theory 62(12), 7475–7490 (2016)
2. Hedayatifar, L., Hassanibesheli, F., Shirazi, A., Farahani, S.V., Jafari, G.: Pseudo paths towards minimum energy states in network dynamics. Physica A: Statistical Mechanics and its Applications 483, 109–116 (2017)
3. Heider, F.: Attitudes and cognitive organization. The Journal of psychology 21(1), 107–112 (1946)
4. Laloux, L., Cizeau, P., Bouchaud, J.P., Potters, M.: Noise dressing of financial correlation matrices. Physical review letters 83(7), 1467 (1999)
5. Ledoit, O., Wolf, M.: Honey, i shrunk the sample covariance matrix. The Journal of Portfolio Management 30(4), 110–119 (2004)
6. Markowitz, H.: Portfolio selection. The Journal of Finance 7(1), 77–91 (1952)
7. Marsili, M., et al.: Dissecting financial markets: sectors and states. Quantitative Finance 2(4), 297–302 (2002)

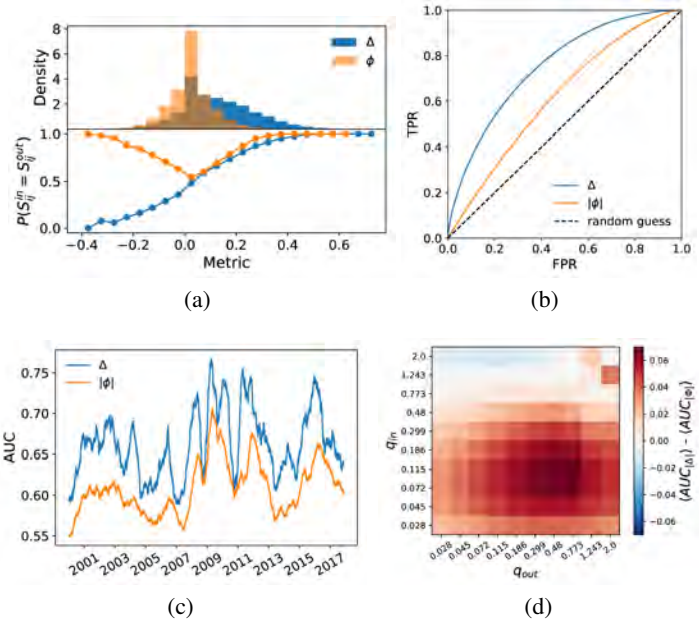


Fig. 1. (a) The lower subplot is the probability to preserve the in-sample sign in the out-of-sample on 2011-18-04 for different values of the discrimination parameter binned in steps of 0.05, and the upper subplot is the related marginal distribution. (b) ROC curve for 2011-18-04 for calibration and test windows of 155 days; (c) evolution of AUC for the two models for calibration and test windows of 155 days; (d) heatmap of the difference between the average AUC of the two discrimination variables Δ and $|\Phi|$ for different calibration and test window lengths.

8. Michaud, R.O.: The markowitz optimization enigma: Is optimizedoptimal? Financial Analysts Journal 45(1), 31–42 (1989)
9. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., Stanley, H.E.: Random matrix approach to cross correlations in financial data. Physical Review E 65(6), 066126 (2002)

Hodge decomposition of Bitcoin money flow among big players

Yoshi Fujiwara[1], Rubaiyat Islam[1] Shinya Kawata[1,2], and Hiwon Yoon[2]

[1] Graduate School of Simulation Studies, University of Hyogo, Kobe 650-0047, Japan,
yoshi.fujiwara@gmail.com,
[2] CMD Laboratory Inc., Tokyo, 151-0051, Japan

Extended Abstract

How money flow among users of Bitcoin is an interesting question in order to understand the dynamics on the complex network of Bitcoin transactions among users. We employ the data of blockchain in the Bitcoin from 2013 to 2018 (compiled by a Hungary research group [3, 2, 1]), utilize a simple algorithm [4] to partially identify anonymous users from addresses, and construct snapshots of temporarily changing network with the users as nodes and the transactions as directed edges. The nodes are selected as "big players" by a certain criterion about the amount of money involving those nodes. In order to understand (1) circular flows and (2) upstream/downstream in the entire network, we use the so-called Hodge decomposition (or Helmholtz-Hodge decomposition) to uncover the structure of (1) and (2). We can find about which big players are located at upstream and downstream side of money flow, how circular flow is present among them, and possibly how such dynamics is related to the exchange market of Bitcoin.

Results: Topology in terms of the so-called "bowtie" structure is given in Fig. 1. The "core" (strongly connected component) is relatively small, with a similar size as the "IN" and "OUT", nodes reachable to and from the core, respectively. Note that the maximum shortest distances from the core to IN and OUT are surprisingly large. Flow in terms of the Hodge potential is given in Fig. 2. Probability distribution has heavy tails at large positive values (upstream) and small negative values (downstream) (see left figure). Also one can observe that each node's potential can quantify the location in the upstream/downstream of the flow (see right figure).

Appendix: method of Hodge decomposition

Consider a directed network with an adjacency matrix A_{ij}, i.e.

$$A_{ij} = \begin{cases} 1 & \text{if there is a directed edge from node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$A_{ii} = 0$ by assumption. Denote the number of nodes by N. Define a flow F_{ij} by

$$F_{ij} = A_{ij} - A_{ji} \tag{2}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

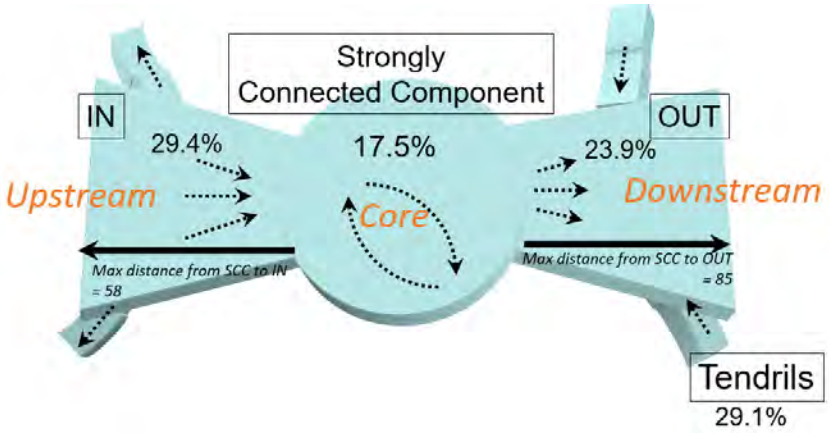


Fig. 1. Daily flow of Bitcoin cryptocurrency (sample of October 10, 2017) in terms of the so-called "bowtie" structure, namely strongly connected componet ("core"), nodes reachable to the core ("IN"), nodes reachable from the core ("OUT"), and the rest of nodes ("Tendrils"). IN and OUT can be regarded as upstream and downstream of the flow.

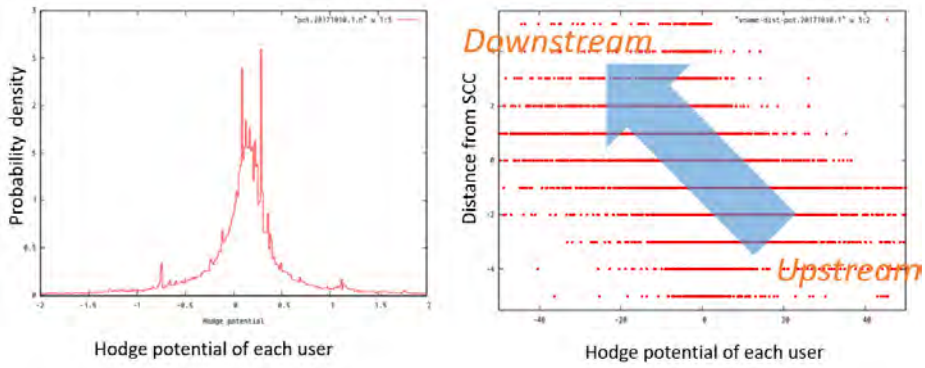


Fig. 2. *Left*: Probability distribution of Hodge potentials for all the nodes. The average of all the potentials is zero by definition. *Right*: Each node's Hodge potential (horizontal axis) and shortest distance from the core (strongly connected component).

and a weight w_{ij} by

$$w_{ij} = A_{ij} + A_{ji} \tag{3}$$

Note that w_{ij} is symmetric:

$$w_{ij} = w_{ji} \tag{4}$$

and non-negative in the sense that

$$w_{ij} \geq 0 \tag{5}$$

for any pair of i and j.

Hodge decomposition is

$$F_{ij} = w_{ij}(\phi_i - \phi_j) + F_{ij}^{(\mathrm{loop})} \tag{6}$$

where ϕ_i is a potential of node i, and $F_{ij}^{(\mathrm{loop})}$ is divergence-free by definition, i.e.

$$\sum_j F_{ij}^{(\mathrm{loop})} = 0 \tag{7}$$

for $i = 1, \ldots, N$. From (6) and (7), given F_{ij} and w_{ij}, one has a linear equation to determine ϕ_i:

$$\sum_j L_{ij}\phi_j = \sum_j F_{ij} \tag{8}$$

for $i = 1, \ldots, N$. Here

$$L_{ij} = \delta_{ij}\sum_k w_{ik} - w_{ij} \tag{9}$$

and δ_{ij} is Kronecker delta:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Obviously, from the symmetry (4), L_{ij} is symmetric:

$$L_{ij} = L_{ji} \tag{11}$$

It is easy to extend the above procedure of Hodge decomposition for a binary directed network to a weighted directed network. We apply the method to the Bitcoin money flow among big players in order to identify which users are located in the upstream or the downstream parts of the network based on each user i's Hodge potential ϕ_i.

References

1. Juháaz, P., Stéger, J., Kondor, D., Vattay, G.: A bayesian approach to identify bitcoin users. PLoS ONE 13(12), e0207000 (2018)
2. Kondor, D., Csabai, I., J., S., Pósfai, M., Vattay, G.: Inferring the interplay between network structure and market effects in bitcoin. New Journal of Physics 16(12), 125003 (2014)
3. Kondor, D., Pósfai, M., Csabai, I., Vattay, G.: Do the rich get richer? an empirical analysis of the bitcoin transaction network. PLoS ONE 9(2), e86197 (2014)
4. Reid, F., Harrigan, M.: An analysis of anonymity in the bitcoin system. In: Altshuler, Y., Elovici, Y., Cremers, A., Aharony, N., Pentland, A. (eds.) Security and Privacy in Social Networks. pp. 197–223. Springer, New York (2013)

Shock Contagion in the World Economy – some results from a correlation study

Tamás Sebestyén[1] and Zita Iloskics[2]

[1] University of Pécs, Faculty of Business and Economics
Pécs, Rákóczi út 80., Hungary,
sebestyent@ktk.pte.hu
[2] University of Pécs, Faculty of Business and Economics
Pécs, Rákóczi út 80., Hungary,

1 Introduction

The realization that the global economic system is, and must be analyzed as, a complex system is becoming common sense ([9], [10]). Many attempts have been recently made to reveal how the microstructure of the economy affects aggregate outcomes, especially in the field of intersectoral transactions ([1],[2]), interregional trade ([14]) or the banking system ([3], [4], [7]). In addition to these efforts, the structure of the global economy as represented by different flows and connections between countries has also been on the agenda for a while. Along this line, the study of international trade networks gained special attention due to the easily available data and the importance of these economic connections ([8], [15]). A vein within this literature argue that overall instability in the complex global economic system and the 2008 crisis in particular is a result of increased globalization and complexity of the underlying notworks ([16], [13], [11]). Apart from investigating trade networks, attention is also directed towards how national economies affect each other at the global level and how shocks propagate through the system of the global economy ([12], [5], [6]). Also, interconnected risks in th eglobal economy and cascades on this risk network has been analyzed recently ([17], [18]).

In this paper we augment the understanding of shock propagation in the global economy using network analysis. We construct a dataset reflecting the network of countries where the ties show estimated contribution of economic changes (as measured by the change in GDP) in one country to changes in the other. Using available longitudinal data we have a dataset covering more than 40 years for 27 countries (allowing for a long-term analysis) and a shorter period comprising the last 20 years for 40 countries (which allows for a before-after analysis of the recent financial crisis). We estimate shock propagation by (i) lagged correlation between GDP growth rates and (ii) estimating Granger-causality on GDP growth rates. The long time coverage allows for a dynamic evaluation of the world economy through the lense of contagion as well as analyzing snapshots around the recent economic crisis in 2008-2009. Using our dataset we try to answer the following questions:

– To what extent did the contagion structure of the world economy change over the past decades? Can we detect a sign of globalization in this respect?

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

– Is the recent crisis a special event with regards to its consequences on the topology of contagion, or does it fit into the pattern of previous turmoils?
– What are the central countries with respect to contagion and does this position correlate with their openness?

2 Results

In order to investigate the previous questions, we used a rolling window of subsamples on our 40 year long panel of GDP growth rates. Every time window consisted of 52 quarters which allows us to estimate Granger-causality (ensuring also stationarity of the within-time-window series). From the estimated causalities we mapped a contagion network where a tie exists if the GDP growth rate of country A Granger causes the GDP growth rate of country B. Rolling these 52 quarters time window through our total sample, we are able to show the evolution of this contagion network over time. Figure 1 shows how the density of this contagion network changed over time, as measured by the ratio of observed links to the possible number of links in the network. It is easy to observe that (i) there is a positive overall trend in the density which underlines the increasingly complex nature of the global economic system and (ii) the recent crisis had a huge effect on the network structure.

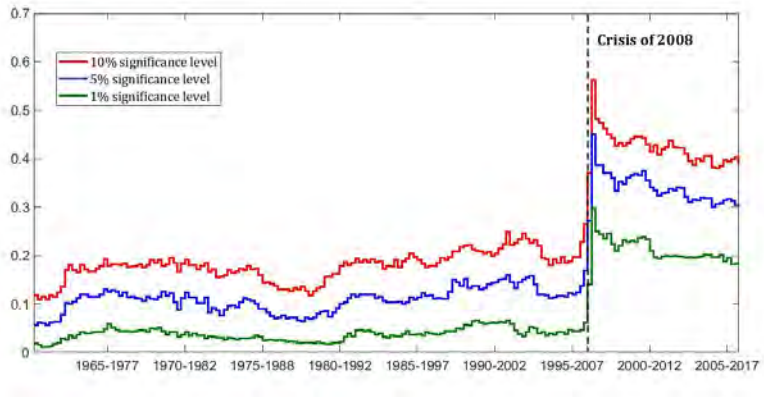


Fig. 1. The evolution of the density (number of observed connections over the possible number of connections) of the global economic contagion network, based on Granger causality. The horizontal axis depicts rolling time windows: one tick corresponds to a time window of 52 quarters on which Granger causality is estimated. Different colors reflect estimations at different significance levels.

Using the dataset with 40 countries in the last 20 years, we run an analysis where we analyze the differences between the before-, under- and after-crisis periods with respect to the recent financial crisis in 2008 and 2009. An inspection of the structure of these snapshot of the contagion network reveals that central countries are different from in

the three periods. It is also interesting that large countries does not necessarily come up as central ones in this network.

Finally, we run an analysis where three centrality measures (degree, betweenness and PageRank) were correlated with a measure of economic openness (share of exports and imports in GDP) of the countries. The results show that there is a clear positive relationship between openness and centrality in the contagion network, i.e. the extent to which countries are transmitters of shocks in the global economy. This pattern was overrun by the turmoil in 2008-2009 rendering the relationship insignificant, but since 2010 the positive correlation is visible again (refer to Figure 2 for scatter plots of Page Rank centrality versus openness in the periods before and during the crisis – after crisis patterns are similar to those before the crisis).

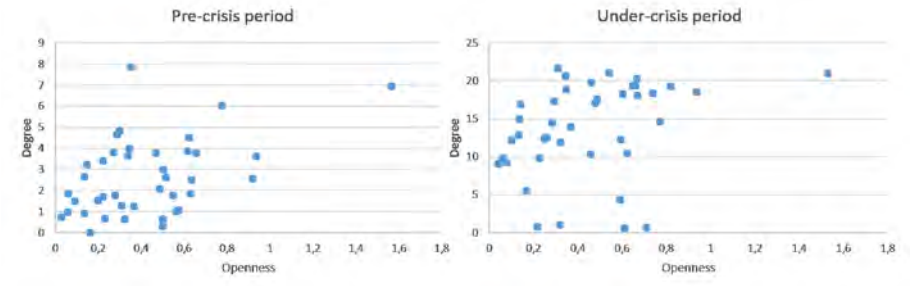


Fig. 2. Correlation between closeness (horizontal axis) and page rank centrality (vertical axis) before and during the 2008-2009 crisis.

Summary. In this paper we used a longitudinal dataset to shed light on the evolution of the contagion network across national economies of the world. Using lagged correlation and Granger-causality test we constructed a dynamic network of ties representing shock contagion across economies. The analysis of this network reveals that the contagion network become increasingly dense over the past decades, evidencing the increasingly complex nature of the global economy. Also, we found that the 2008-2009 crisis had an outstanding effect on this network by increasing its density. Finally, we have shown that centrality within this contagion network correlates with economic openness.

References

1. Acemoglu, D., Carvalho, V.M., Ozdaglar, A., Tahbaz-Salehi, A. (2012): The network origins of aggregate fluctuations. Econometrica, 80(5), 19772016.
2. Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A. (2015): Networks and the macroeconomy: an empirical exploration, Bank of Finland Research Discussion Papers 25.
3. Allen, F., Gale, D. (2000): Financial contagion. Journal of Political Economy, 108(1), 133.
4. Allen, F., Babus, A., Carletti, E. (2010): Financial connections and systemic risk, NBER Working Papers 16177.

5. Askari, M., Shirazi, H., Samani, K.A. (2018): Dynamics of financial crises in the world trade network. Physica A: Statistical Mechanics and its Applications, Volume 501, 1 July 2018, Pages 164-169
6. Blonigen, B.A., Piger, J., Sly, N. (2012): Comovement in GDP trends and cycles among trading patners. Journal of International Economics, Volume 94, Issue 2, November 2014, Pages 239-247
7. Elliott, M., Golub, B., Jackson, M.O. (2014): Financial networks and contagion. American Economic Review, 104(10), 31153153.
8. Fagiolo, G., Reyes, J., Schiavo, S. (2009): World-trade web: Topological properties, dynamics, and evolution, Physical Review E, Vol. 79, Iss. 3
9. Farmer, J.D., Foley, D. (2009): The economy needs agent-based modelling. Nature, 460, 685686.
10. Farmer, J.D. (2013): Economics needs to treat the economy as a complex system. CRISIS publications working paper.
11. He, J., Deem, M.W. (2010): Structure and Response in the World Trade Network, Phisical Review Letters, 105, 198701
12. Sander, H., Kleimeier, S. (2003): Contagion and causality: an empirical investigation of four Asian crisis episodes. Journal of International Financial Markets, Institutions and Money, Volume 13, Issue 2, April 2003, Pages 171-186
13. Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., White, R.D. (2009): Economic Networks: The New Challenges. Science, Vol 325., pp. 422425
14. Sebestyén, T. (2017): Moving beyond the iceberg model. Economic Modelling, 67, pp. 159-174.
15. Serrano, Á., Boguñá, M. (2003): Topology of the world trade web, Physical Review E, Vol. 68, Iss. 1
16. Sheng A. (2010): Globalization and Growth Implications for a Post-Crisis World, Financial crisis and global governance: A network analysis, eds Spence M, Leipziger D (World Bank, Washington, DC), pp 6993
17. Szymanski, B.K., Lin, X., Asztalos. A., Sreenivasan. S. (2015): Failure dynamics of the global risk network. Scientific Reports, 5, an. 10998.
18. Xin, L., Moussawi, A., Korniss, Gy., Bakdash, J.Z., Szymanski, B.K. (2017): Limits of Risk Predictability in a Cascading Alternating Renewal Process Mode. Scientific Reports, 7, an. 6699.

Measurement of Value of Firms Based on Their Stock Ownership Relations

Haruka Kato[1], Hitomi Sato[1], Yuichi Kichikawa[2], Hiroshi Iyetomi[2,5], Wataru Souma[3,5], and Tsutomu Watanabe[4,5]

[1] Graduate School of Science and Technology, Niigata University, Niigata 950-2181, Japan,
[2] Faculty of Science, Niigata University, Niigata 950-2181, Japan
[3] College of Science and Technology, Nihon University, Funabashi 274-8501, Japan
[4] Graduate School of Economics, The University of Tokyo, Tokyo 113-8656, Japan
[5] The Canon Institute for Global Studies, Tokyo 100-6511, Japan

1 Introduction

Stock cross-holding was used to prevail over listed firms in Japan. Such a conventional capital transaction method is fading away under the governmental guidance. However, the corporate ownership is still so complicated because listed firms are interconnected through holding shares one another in a multilateral way. The objective of this study is to shed empirical light on the "Who possesses whom" structure making full use of the latest methods in network science.

2 Ownership Network

Here we use the major shareholder database compiled by Toyo Keizai Inc. [1] during the period from 1985 to 2009. The proprietary database provides us with the top 30 (20 before 2002) shareholders' information for all listed firms in Japan. From the database we construct a directed network each year in which firms and shareholders constitute nodes and ownership relations from shareholders to firms, directed links with the proportionate fractions of the market capitalization value of firms as their weights. Note that the results shown below are those obtained for the data in 2008.

Figure 1 depicts the bow-tie structure [2] of the Japanese ownership network in 2008. The network has a giant strongly connected component (GSCC) to which about one third of listed firms belong. Any pairs of nodes in the GSCC are connected in both ways, that is, the two nodes are

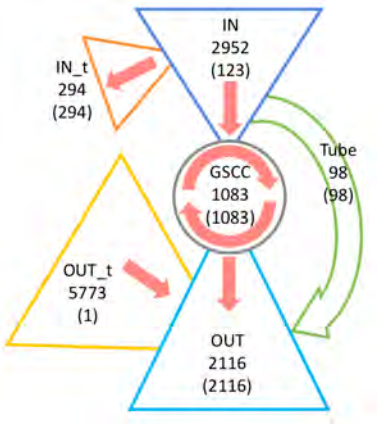


Fig. 1. Bow-tie diagram of the Japanese ownership network in 2009. The figure associated with the designation of each component is the number of nodes in it and the figure in the parentheses, the number of listed firms among them.

on a loop of ownership. The GSCC thus represents multilateral cross-holding, making the ownership relations obscure.

To resolve the fundamental issue stemming from the existence of the GSCC, we invoke on the Helmholtz-Hodge decomposition [3]. It uniquely breaks up the flow structure on a directed network into gradient and circular flow components. The gradient flow component illuminates the hierarchical structure in the GSCC; the multilateral cross-holding embedded in it is totally canceled out by the circular flow component. Although a number of works have already been carried out to elucidate the stock ownership structure at a country level and on a global scale, the use of the Helmholtz-Hodge decomposition distinguishes this work from the previous ones.

3 Ultimate Owners of Firms

Once we replace the flow structure in the GSCC with its gradient flow component obtained by the Helmholtz-Hodge decomposition, we can trace the stock holding relations of a given listed firm back to its ultimate owners. The procedure is schematically shown in Fig. 2, where the target firm, numbered as 7, is assumed to be worth 100 in some monetary units and its value is ultimately ascribed to firm 1, firm 2, shareholder of firm 3, and shareholder of firm 5 according to the proportional division.

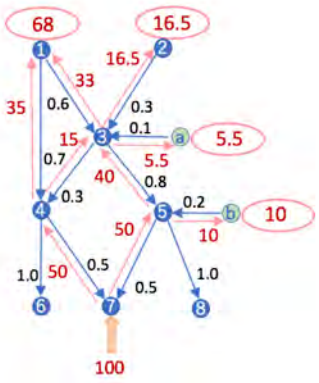


Fig. 2. Demonstrative hierarchical network describing how to compute the ultimate ownership.

To characterize the shareholders' distribution, we divide the nodes in the ownership networks into the following seven categories: 1) Japanese listed firms in financial sector, 2) Japanese listed firms other than in financial sector, 3) the rest of Japanese firms, 4) overseas firms, 5) employee stock ownership, 6) executives, and 7) others. We then quantify the difference between the distribution, $\boldsymbol{p} = (p_1, \cdots, p_7)$ with $p_1 + \cdots + p_7 = 1$ $(0 \leqslant p_k \leqslant 1)$, of the original owners and the distribution $\boldsymbol{q}$ of the corresponding ultimate owners for listed firms using the normalized L2 distance defined by

$$d(\boldsymbol{p}, \boldsymbol{q}) = \frac{\|\boldsymbol{p} - \boldsymbol{q}\|^2}{\|\boldsymbol{p}\|^2 + \|\boldsymbol{q}\|^2} \leqslant 1, \tag{1}$$

where $d(\boldsymbol{p}, \boldsymbol{q})$ takes its maximum when the two vectors are orthogonal.

The results are summarized in Fig. 3. Here the median of $d(\boldsymbol{p}, \boldsymbol{q})$ is listed for each industry as a representative value of the difference between the two distributions to get rid of outliners. We see that the constitution of the ultimate owners in industries such as shipping, steel, transportation equipment, mining, and air transport are significantly different from that of the primary owners. This is because firms in those industries are mainly located on downstream side in the ownership hierarchy.

4 Hierarchical Market Value

In the course of computing the ultimate ownership, we have devised a new measure for evaluating firms from the market values. We recall Fig. 2. For instance, firm 3 can be regarded as a partial owner of firms 4 through 8, which are firms sitting at lower positions in the hierarchical ownership tree. We coin *hierarchical market value* to refer to such hierarchically accumulated financial assets of a firm. Figure 4 compares the hierarchical market values for listed firms with their actual market values calculated from the current stock prices and the total numbers of issued shares. Remarkably, we find a non-negligible number of extraordinary firms whose hierarchical market values are comparable to or even higher than the corresponding actual market values. Those may be hidden firms for which the market overlook their additional value yielded through the complicated ownership network, or large firms suffering from *conglomerate discount*.

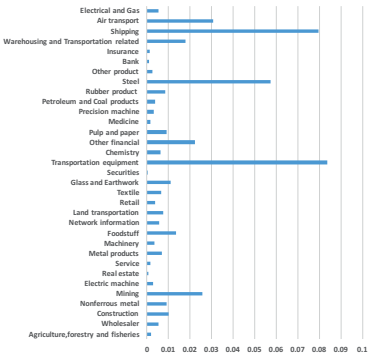


Fig. 3. The normalized L2 distances between the distribution of original owners and that of the ultimate owners for listed firms, where the medians within industries are shown.

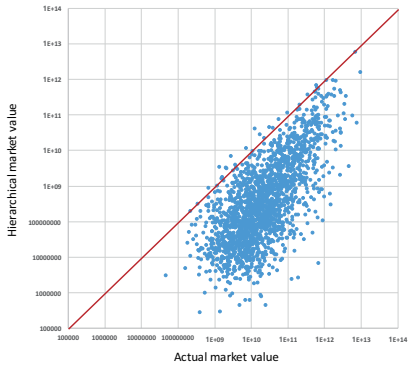


Fig. 4. Comparison between the actual and the hierarchical market values for listed firms with corresponding ultimate owners for listed firms, the diagonal line on which the two market values are identical.

Summary. We have analyzed the stock ownership relations in Japan from a network theoretic point of view. The Helmholtz-Hodge decomposition eliminates the circular ownership relations to allow us to determine the ultimate owners of listed firms. Also we have devised a new measure to evaluate value of firms by taking advantage of the computation of the ultimate ownership. This work was supported by Nomura Foundation and JSPS KAKENHI (17KT0034, 18K03451).

References

1. https://biz.toyokeizai.net/en/data/service/detail/id=862
2. Baldi, P., Frasconi, P., Smyth, P.: Modeling the Internet and the Web: Probabilistic methods and algorithms. John Wiley & Sons (2003)
3. Jiang, X., Lim, L.H., Yao, Y., Ye, Y.: Statistical ranking and combinatorial hodge theory. Mathematical Programming 127(1), 203–244 (2011)

Fire sales as multistate contagion on bipartite networks

Tomokatsu Onaga,[1] Fabio Caccioli,[2] Teruyoshi Kobayashi[3,*]

[1] The Frontier Research Institute for Interdisciplinary Sciences and Graduate School of Information Sciences, Tohoku University, Sendai, Japan
[2] Department of Computer Science, University College London, London, UK
[3] Department of Economics and Center for Computational Social Science, Kobe University, Kobe, Japan
*kobayashi@econ.kobe-u.ac.jp

1 Introduction

Fire sales in financial markets play a crucial role in initiating and deepening financial crisis. If an important financial institution (hereafter bank), like Lehman Brothers, collapsed, the financial assets held by the failed bank would be sold (i.e., liquidated) in the financial market to repay its debts to the creditors. The sales of financial assets cause the prices of the assets to go down, which could in turn lead other banks to default because of the capital losses they incur through overlapping portfolios. This feedback between bank defaults and decline in asset prices can be considered as cascading failures on a bipartite network where nodes are banks and financial assets and edges represent the asset holdings (i.e., portfolios) of banks (Fig. 1(a)).

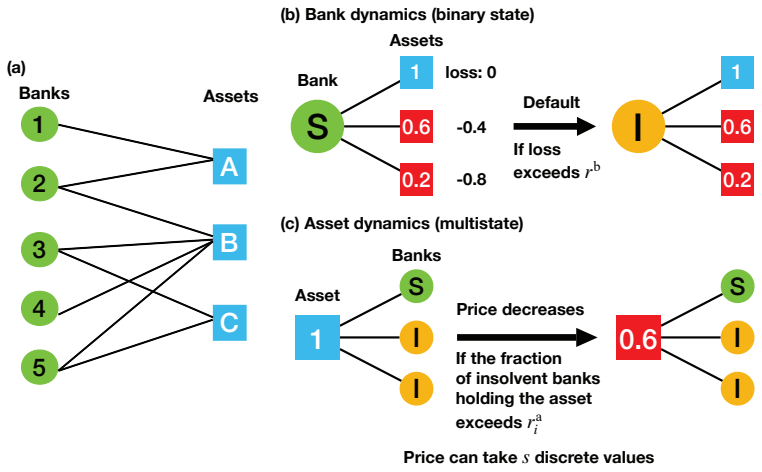


Fig. 1. Schematic of fire sales. (a) Bipartite network of banks and financial assets. Edges represent the asset holding of banks. (b) Dynamics of binary bank state. S and I stand for solvent and insolvent, respectively. (c) Dynamics of asset price. Each asset takes one of the s discrete values (i.e., asset prices), depending on the fraction of defaulted banks among all the banks that hold the asset.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Over the past decade, many attempts have been made to characterize fire sales by simulating a threshold model on a bipartite network [1–3]. However, unlike the standard threshold models [4], it is difficult to solve the model of fire sales in an analytical manner for two reasons; First, in practice, the prices of assets can take any positive value. The multistate property of asset prices prohibits us from exploiting the solution methods that have been developed for the standard binary-state cascade models [4, 5]. Second, there are two types of nodes, banks and assets, whose possible states are intrinsically different. While asset prices can take continuous values, the state of each bank is binary, namely solvent or insolvent. Due to these difficulties, previous studies on fire sales based on a model of asset-bank bipartite network rely on numerical simulations [1, 3, 6].

Here, we offer a model of multistate contagion to study fire sales, namely the interaction between bank defaults and asset price declines, in an analytical manner. To do so, we attempt to capture the continuity of prices as a limit of multistate cascades on bipartite networks [7, 8].

2 Model

In the model, each bank takes one of the two states: solvent (S) or insolvent (I). A bank changes its state from solvent to insolvent if the total loss exceeds r^{b} (Fig.1(b)). Asset price p takes one of the s values: $p \in \{1,(s-1)/s,(s-2)/s,\ldots,1/s\}$. An asset price becomes $1-i/s$ if the fraction of insolvent banks among all the banks that hold the asset exceeds r_i^{a} (see, Fig. 1(c)), where

$$r_i^{\mathrm{a}} = 1-\left[1-\left(\frac{i}{s}\right)^{\alpha}\right]^{\frac{1}{\alpha}}, \quad i=0,1,\ldots,s-1. \tag{1}$$

α is a parameter that modulates the elasticity of asset prices.

3 Results

To investigate the model analytically, we extend the theory of multistate dynamical processes [7, 8] to a model of complex contagion with heterogeneous states. Specifically, we use an approximation technique based on the approximate master equation (AME) [8] to simultaneously calculate both the expected fraction of banks that are defaulted due to cascading declines in asset prices and the shares of assets in state $i=0,\ldots,s-1$.

The average final fraction of insolvent banks calculated by our method is shown in Fig. 2. The figure reveals that our theory matches the numerical result fairly well. By increasing the mean number of assets held by a bank, z, the possibility of global default cascades begins to rise continuously at $z \approx 2.8$. By increasing z further, the possibility of global cascades then disappears in a discontinuous manner at $z \approx 4.5$. Our theory well captures these two different kinds of transitions.

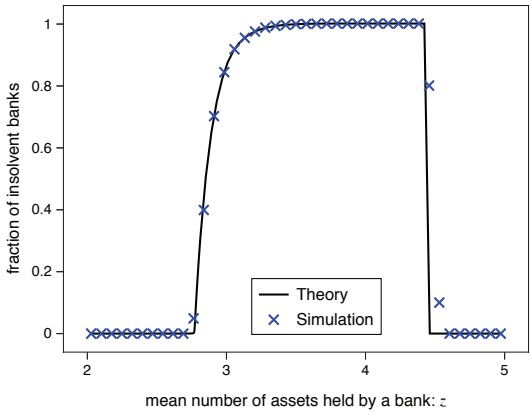


Fig. 2. Theory and Monte Carlo simulations for the configuration model with $N = 4000$ banks and $M = 8000$ assets. Crosses denote the size of global cascades (i.e., fraction of insolvent banks) averaged over 10 numerical simulations, conditional on more than 5% of banks being insolvent. The solid line represents the fraction of insolvent banks in theory. We set the number of asset states at $s = 5$, the elasticity of asset price at $\alpha = 0.8$, and the threshold of bank default at $r^{\mathrm{b}} = 0.15$. At $t = 0$, a fraction $\rho_0 = 0.001$ of banks are insolvent and all assets are in state 0 (i.e., $p = 1$).

4 Summary

In this work, we offer a model of multistate contagion to study the interaction between bank defaults and asset price declines in an analytical manner. We present an analytical approach by extending the theory of multistate cascades [7, 8] to a model of complex contagion with heterogeneous states. The advantage of our approach is twofold: First, as the number of states for asset prices increases, the analytical result can better describe continuous prices. Our method could therefore save a considerable computational effort to calculate the true cascade size. Second, our model allows to derive analytical cascade conditions with which we can identify the parameter region that would lead to global cascades.

References

1. Huang, X., Vodenska, I., Havlin, S., Stanley, H.E.: Scientific Reports **3** (2013) 1219.
2. Caccioli, F., Shrestha, M., Moore, C., Farmer, J.D.: Journal of Banking and Finance **46** (2014) 233–245.
3. Caccioli, F., Barucca, P., Kobayashi, T.: Journal of Computational Social Science **1** (2018) 81–114.
4. Gleeson, J., Cahalane, D.: Phys. Rev. E **75** (2007) 56103.
5. Hurd, T.R.: *Contagion! Systemic Risk in Financial Networks* (2016) Springer.
6. Levy-Carciente, S., Kenett, D.Y., Avakian, A., Stanley, H.E., Havlin, S.: Journal of Banking & Finance **59** (2015) 164–181
7. Melnik, S., Ward, J.A., Gleeson, J.P., Porter, M.A.: Chaos **23** (2013) 013124.
8. Fennell, P.G., Gleeson, J.P.: SIAM Review **61** (2019) 92–118.

A network approach to the analysis of bitcoin

Alexandre Bovet[1,2], Carlo Campajola[3], Francesco Mottes[4], Valerio Restocchi[5,6], Nicolò Vallarano[7], Tiziano Squartini[7], and Claudio J. Tessone[8]

[1] ICTEAM, Universitè Catholique de Louvain, B-1348 Louvain-la-Neuve (Belgium)
[2] naXys, University of Namur, B-5000 Namur (Belgium)
[3] Scuola Normale Superiore, I-56126 Pisa (Italy)
[4] Physics Department, Università di Torino, I-10125 Torino and Istituto Nazionale di Fisica Nucleare (Italy)
[5] University of Southampton, SO171BJ Southampton (UK)
[6] The University of Edinburgh, EH89YL Edinburgh (UK)
[7] IMT School for Advanced Studies Lucca, I-55100 Lucca (Italy)
[8] URPP Social Networks and UZH Blockchain Center, University of Zurich, CH-8050 Zürich (Switzerland)

1 Introduction

Cryptocurrencies are distributed systems that allow exchanges of native tokens among participants [1]. The availability of their complete historical bookkeeping opens up the possibility of understanding the relationship between aggregated users' behavior and the cryptocurrency pricing in exchange markets. Here we analyze the properties of the transaction network of Bitcoin over a period of nine years since the Bitcoin creation and involving 16 million users and 283 million transactions. To this aim, we consider both the User Network and the Address Network representations at two different time scales (i.e. daily and weekly). *Addresses* are pseudo-anonymous alpha-numeric strings which are public signs on every transactions to establish bitcoin ownership, *users* are clusters of addresses produced after heuristic rules inferred by the bitcoin protocol[2]. By analyzing these networks, we show the existence of causal relationships between Bitcoin price movements and changes of its transaction network topology.

2 Results

Stylized facts While the number of nodes and links increases with time, link density has a steady decrease: regardless of the considered representation the networks become sparser over time, as a consequence of the the average degrees being tightly bounded around 10 over the whole period considered.

The degree distributions are heavy-, right-tailed: a vast majority of nodes with low-degree coexists with few hubs, in line with the disassortative empirical evidences of all representations. We investigate the power-law nature of the networks by employing a double Kolmogorov-Smirnov statistical test [4]. The out-degree distribution bore the most interesting results: the power-law hypothesis cannot be refused (with a .05 confidence) in 54% cases before 2014 and drops to 26% afterwards.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The two regimes Studying the standard deviation of the degree distributions evolution over time, two different regimes emerge before and after 2014. In the first period the degree distribution tends to be sharper during price surges and wider during the price drawdowns, while after 2014 all moments tend to become larger in value. Our interpretation is that in Bitcoin beginnings a sort of herding mechanism was identifiable: when many users behave similarly (low standard deviation and skewness) price increases; when the price peak is crossed (i.e. a drawdown starts) individuals connectivity gets more heterogeneous, widening the degree distribution. After 2014 the situation gets less readable, and it's harder to relate the moments to the prices : 2014 stands out as a natural threshold year for in February the in-famous Mt.Gox happened, triggering a huge price drop and fostering a dramatic change in Bitcoin ecosystem.

Reciprocity seems to evolve in a similar way: until 2014 it is highly volatile, oscillating between 0.01 and 0.05 and showing an interesting overlap with the Bitcoin price bubbles identified in [5]. After 2014 it revolves around a 0.02 slowly descending to an all-time low of less than 0.01 around the end of the well known 2017 bubble.

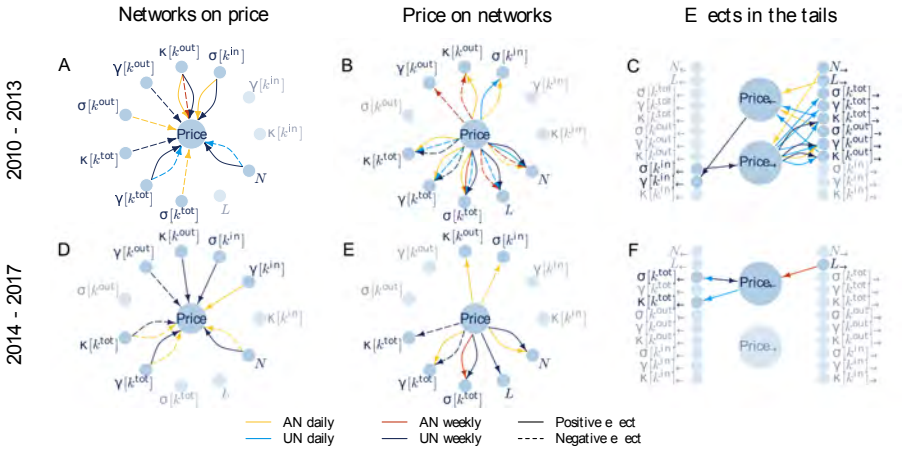


Fig. 1. The Granger causality tests results. From the left: effects in mean of network variables on price returns (**A**,**D**), effects in mean of price returns on network variables (**B**,**E**), mutual effects in tail between network variables and price returns (**C**,**F**). The tests are performed on the four network representations, where *AN* stands for *addresses network* and *UN* for *users network*. Thickness of the lines shows the magnitude of the effect.y

Causality analysis We performed a Granger causality test[6] over the two sub-samples induced by the two degree distributions regimes (before and after 2014), in order to investigate the mutual influence of network variables and price dynamic(represented by the log return of the BTC-US dollar price).

As a result we identify two different feedback-loops in action. On a weekly scale the number of nodes growth leads to a increase in the Bitcoin price, which in reverse leads to an increase in the number of nodes, in a slow feedback mechanism over several

weeks. On the other hand, an increase in the out-degree kurtosis leads to a price increase on the weekly timescale, which by the first loop causes the number of nodes to increase. Observing that the number of nodes and the out-kurtosis are positively correlated, we are able to close the second loop, which although being weaker, has interesting consequences. By the Granger tail analysis[7] we find out that abnormal sudden increase in the out-kurtosis actually causes price abrupt decrease: a possible explanation for the frequent crashes of 2014 is given by the second feedback loop reaching an unsustainable growth rate, leading to violent mean reversions in the market.

3 Summary

Our results distinguish two different phases in the evolution of the Bitcoin ecosystem. Especially in the first period, we detect topological hints of an herding mechanism driving BTC-US dollar price in a repetitive course of steady surges and fast crashes.

Focusing on the growth mechanisms, we identified feedback-loops between BTC-US dollar price and some structural properties of the Bitcoin network.

References

1. Antonopoulos, A.M., 2017. Mastering Bitcoin: Programming the open blockchain. " O'Reilly Media, Inc.".
2. Harrigan, M. and Fretter, C., 2016, July. The unreasonable effectiveness of address clustering. (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld) (pp. 368-373). IEEE.
3. Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G.M. and Savage, S., 2013, October. A fistful of bitcoins: characterizing payments among men with no names. In Proceedings of the 2013 conference on Internet measurement conference (pp. 127-140). ACM.
4. Restocchi, V., McGroarty, F. and Gerding, E., 2019. The stylized facts of prediction markets: Analysis of price changes. Physica A: Statistical Mechanics and its Applications, 515, pp.159-170.
5. Gerlach, J.C., Demos, G. and Sornette, D., 2019. Dissection of Bitcoins multiscale bubble history from January 2012 to February 2018. Royal Society open science, 6(7), p.180643.
6. Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, pp.424-438.
7. Hong, Y., Liu, Y. and Wang, S., 2009. Granger causality in risk and detection of extreme risk spillover between financial markets. Journal of Econometrics, 150(2), pp.271-287.
8. Bovet, A., Campajola, C., Mottes, F., Restocchi, V., Vallarano, N., Squartini, T. and Tessone, C.J., 2019. The evolving liaisons between the transaction networks of Bitcoin and its price dynamics. arXiv preprint arXiv:1907.03577.

Who Possesses Whom from a Point of View of the Global Ownership Network

Yuichi Kichikawa[1], Hiroshi Iyetomi[1,4], Yuichi Ikeda[2], and Takayuki Mizuno[3,4]

[1] Faculty of Science, Niigata University, Niigata 950-2181, Japan,
[2] Graduate School of Advanced Integrated Studies in Human Survivability, Kyoto University, Kyoto 606-8306, Japan
[3] National Institute of Informatics, Tokyo 101-8430, Japan
[4] The Canon Institute for Global Studies, Tokyo 100-6511, Japan

1 Introduction

Value of firm is a monetary representation of how well a firm is doing its business, and is recognized as one of the important indicators in management judgment of firms and investors. There are several ways to express value of firm, one of which is stock price. The stock price is considered to reflect the future profitability of the firm and the value of the assets held. Since firm assets include stock assets, if firm B has shares in firm A, it can be considered that the shareholders of firm B will indirectly own part of value of firm A. Such a relationship is actually a very complex network, and it is not obvious who is the true holder of corporate value. So, we attempt to clarify the structure of stock ownership by empirically analyzing the network constructed from Thomson Reuters global equity ownership data, which is a proprietary database daily financial transactions of firms over the world.

2 Data and Methods

From Thomson Reuters global equity ownership data in 2017, we constructed the stock ownership network with firms and shareholders as nodes and stock ownership relationships as links. The direction of the link coincides with the direction of ownership (from owner to stock issuer), and the weight is the current price of shares of the owner. We deal with only the largest connected component (96.7%). The network constructed in this way is a weighted directed network with 203,182 nodes and 4,295,669 links. To elucidate flow structure in the network, we begin with a bow-tie decomposition of the network as has been widely used to understand the structure of various complex networks. The decomposition classifies nodes in a directed network according to the way in which they are mutually connected. Table 1 shows the results of the bow-tie decomposition. Basically the network is hierarchical as it should be. However, we note that there exists a giant strongly-connected component (GSCC). Although the GSCC constitutes only 1.6% of the whole network, it occupies the central core of the network as shown in Fig. 1 below. Furthermore, the GSCC, in which any pairs of nodes are connected in both ways, contains loop flows and hence gives rise to complexity in the ownership relations.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

To unfold such entangled ownership relations, we adopt the Helmholtz-Hodge decomposition [1, 2] which decompose the network flow into hierarchical and circular flow components. The Helmholtz-Hodge potential of nodes in a directed network identifies their hierarchical positions in the flow structure. In contrast, the circular flow component illuminates feedback loops built in the system.

We estimate the indirect ownership relationship by the way as follows. Suppose the firm A has 50% of the firm B's stock, and the firm B has 50% of the firm C's stock. At this time, it is considered that the firm A holds 50% of the value of firm C owned by the firm B, and finally holds 25% of the value of firm C. In this way, it is traced which shareholder finally holds the value of the firm. We call the owner who finally owns the corporate value of a company ultimate owner of that company. However, since the calculation may not converge if the network contains loops, we use Helmholtz-Hodge decomposition for removing the loops. Removing loop flow components obtained by Helmholtz-Hodge decomposition makes the network completely hierarchical.

Table 1. Bow-tie components

IN	43,704 (21.5%)
IN tendril	5,716 (2.8%)
GSCC	3,168 (1.6%)
OUT tendril	96,791 (47.6%)
OUT	9,163 (4.5%)
Tube	2,701 (1.3%)
Other	41,939 (20.7%)
Total	203,182 (100%)

3 Results and Discussion

Figure 1 shows the whole network and GSCC visualized in 3D space. The GSCC is mainly occupied by nodes belonging to Japan and the United States (Japan 69.4%, the United States 21.3%), and the United States stands hierarchically higher position among these countries. The reason is that there are 17,083 links from the United States to Japan in the GSCC, whereas there are only 42 links from Japan to the United States. In addition, there is a difference of about 100 times in the total weight. This kind of hierarchical position bias is not limited to Japan and the United States. Figure 2 shows the comparison of direct holdings and ultimate holdings in G20 countries. The ultimate holdings increase as compared to the direct holdings for countries such as Italy, Germany and Canada, and vice versa for Japan, South Africa and Argentina. These results show that the ultimate owners, determined by taking into around the indirect effects of share ownership, are significantly different from the direct owners even at the country level, and provide a new perspective on corporate value. In some countries,

This work was supported by Nomura Foundation and JSPS KAKENHI (17KT0034, 18K03451).

References

1. Bhatia, H., Norgard, G., Pascucci, V., Bremer, P.T.: The Helmholtz-Hodge decompositiona survey. IEEE Transactions on visualization and computer graphics 19(8), 13861404 (2013)
2. Jiang, X., Lim, L.H., Yao, Y., Ye, Y.: Statistical ranking and combinatorial Hodge theory. Mathematical Programming 127(1), 203244 (2011)

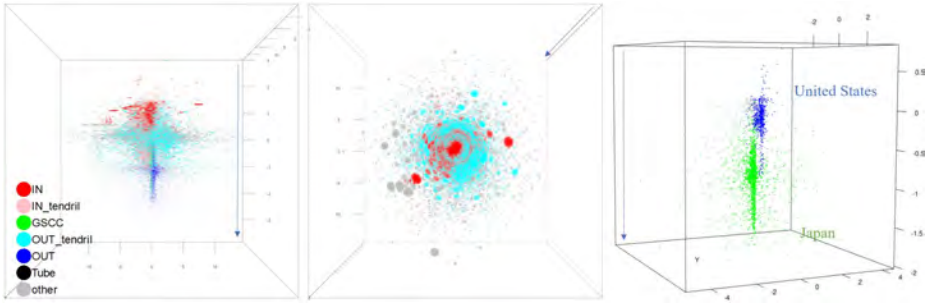


Fig. 1. Visualized in three-dimensional space with different points of view. Nodes are aligned in the z direction according to their values of the Helmholtz-Hodge potential; basically, flows from top to bottom. On the other hand, the x and y coordinates of nodes are determined by the energy minimum principle with a spring-electric model. Left and middle panels are whole network, and right panel shows only GSCC.

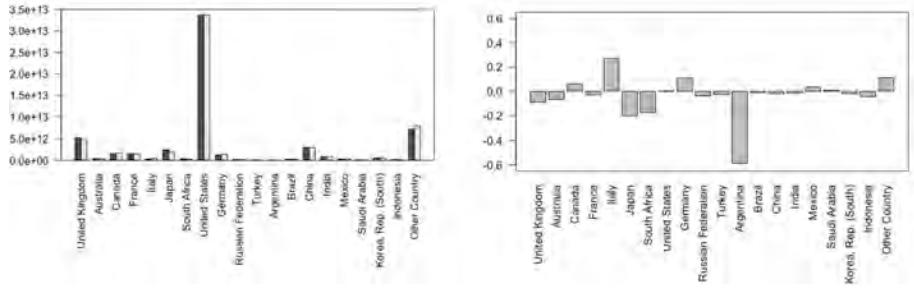


Fig. 2. Comparison between ultimate and direct owner. The left panel shows the amount owned of the direct (black) and ultimate owner (ash) for each G20 country. The right panel shows the rate of increase of amount owned of ultimate owner compared to the direct owner for each G20 country.

Firms' Complexity: Technological Coherence, Performance, and Forecasting

Andrea Zaccaria[1], Louis Barlascini[2] Lorenzo Napolitano[3], Emanuele Pugliese[1], and Luciano Pietronero[2]

[1] Institute for Complex Systems, CNR, Rome, Italy,
and.zaccaria@gmail.com,
[2] Physics Department, "Sapienza" University of Rome, Italy
[3] Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy

The aim of this work [1, 2] is to study firms' technological portfolios using tools borrowed from complexity and network science. In particular, we analyze the relationship between the *coherence* of such portfolios and firms' performance, and we show that such tools can be also used to make specific predictions about firms' patenting activity.
The first issue we investigate is whether the accumulation of knowledge and capabilities related to a coherent set of technologies leads firms to experience advantages in terms of productive efficiency. We analyzed both the balance sheets and the patenting activities of about 70 thousand firms that have filed at least one patent over the period 2004-2013. Each patent contains a number of codes that identify the respective technological sectors. We then define a *network of technological sectors*, whose links are obtained by counting the co-occurrences of the respective codes in companies' portfolios, and then normalizing such countings in a suitable way. The general idea is that two nodes will be close if many companies are patenting in both sectors, and so co-occurrences are used as proxies of the presence of common capabilities.
The network structure can be used to study firms' patenting strategy. More specifically, we introduce firms' *coherent diversification*, a quantitative assessment of each technological portfolio that counts the fields in which a firm is active weighting them with their coherence with respect to the firms global knowledge base (see Figure 1). Such a measure favors companies that patent following blocks of closely related fields with respect to companies with the same breadth of scope, but more scattered, or incoherent, portfolios. In this sense, coherent diversification contains both a diversification element, being correlated with the simple counting of sectors, and a specialization element, being based on the idea that technological portfolios should focus on related sectors (those that are close to each other in the network).
We find that our measure of coherent diversification is quantitatively related to economic performance. In particular, we prove on a statistical basis that it explains *labor productivity* better than standard diversification.
As a second step, we study the time evolution of firms' technological portfolio. We show that by leveraging the concept of coherence we are able to *forecast* the technological sectors in which a specific company will patent in the next years. Moreover, studying specific data regarding merging and acquisitions, we show that companies are willing to pay more for those acquisitions that include technological fields that are coherent with respect to their actual patenting activity.
In conclusion, there is empirical evidence that this evaluation of technological portfolios

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

captures relevant information about the productive efficiency and the future patenting activity of firms. As a consequence, it can be also used to investigate possible synergies within firms and to recommend viable partners for mergers and acquisitions.

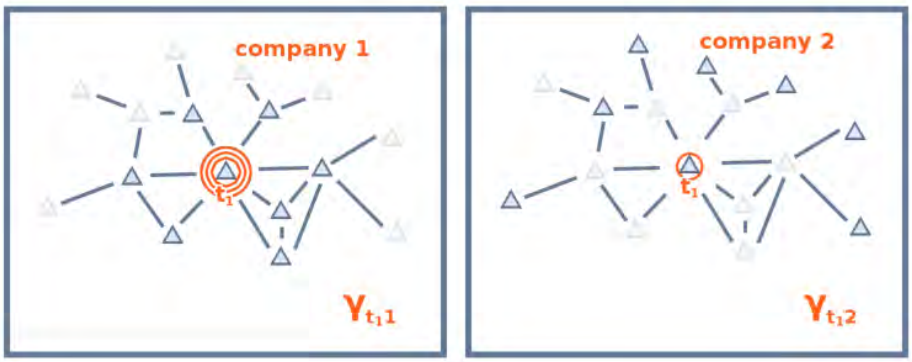


Fig. 1. Schematic illustration of how the intra-firm coherence of technologies may change when two different companies (one on the left, one on the right) are considered. The opaque triangles indicate technological fields in which the two firms hold patents. The two have the same technological diversification, however, the fields in which firm 1 is active are all close in the network, while the portfolio of firm 2 is scattered through the graph. As a consequence, the same technology t_1 is *coherent* with the portfolio of firm 1 (on the left), and not firm 2 (on the right). Such increase of coherence is visualized using the concentric orange circles.

References

1. Pugliese E, Napolitano L, Zaccaria A, Pietronero L (2019) Coherent diversification in corporate technological portfolios. PLOS ONE 14(10): e0223403.
2. Barlascini L, Napolitano L, Zaccaria A, and Pietronero L, in preparation.

Part XV

Political Networks

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Network Analysis of Brexit-Related votes in the 57^{th} UK Parliament

Carla Intal[1] and Taha Yasseri[1,2]

[1] Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK,
[2] Alan Turing Institute, London NW1 2DB, UK,
taha.yasseri@oii.ox.ac.uk

Summary. The British party system is known for its discipline and cohesion, but it remains wedged on one issue: European integration. This was observed both in the days of the EEC in the 1970s and the EU-Maastricht treaty in the 1990s; This work aims to investigate whether this holds true in the Brexit era. We utilise social network analysis to unpack the patterns of dissent and rebellion among pairs of MPs. Using data from Hansard, we compute similarity scores between pairs of MPs from June 2017 until April 2019 and visualise them in a force-directed network. Comparing Brexit- and non-Brexit divisions, we analyse whether patterns of voting similarity and polarity differ among pairs of MPs. Our results show that Brexit causes a wedge in party politics, consistent to what is observed in history.

The British party system is arguably one of the most successful in the world, and many scholars consider the party discipline in the House of Commons as a model that many Governments should follow [1]. Throughout its contemporary history, the strong party values and ideologies that define its two main parties —Labour and Conservative—has lent credibility to the Parliamentary process, setting the landscape for the effective implementation of policies in the British government. It is notable, however, that the cohesion and unity in the modern British party system is persistently wedged by one issue, which is that of European integration [2, 3]. We analyse the voting records of the MPs during the 57th parliment and quantify within party conflicts and cross-party alliences based on voting similarity score (see Figure 1).

We show that Brexit is consistent with the other issues of European integration in the past, in that it creates a wedge in Parliament. We report that there is a strong disparity in MP voting on Brexit divisions compared to non-Brexit. The network analysis showed that while there are two distinct (ideology) clusters on both the Brexit and the non-Brexit case, the inter-connectivities across these clusters differ significantly. In non-Brexit divisions, it is almost certain that MPs follow the party rhetoric, and defying the party whip is largely negligible. As demonstrated by the network visualisation, most cross-party alliances happen within one cluster only and rarely does it ever cross to the other side. Meanwhile, within-party conflicts are also very minimal. On the other hand, in the Brexit divisions, there was a visible blurring of the party line, and cross-cluster interaction is obvious and apparent. There exists strong repulsion across various node pairs, and while in the non-Brexit case, cross-party alliances only happen within

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

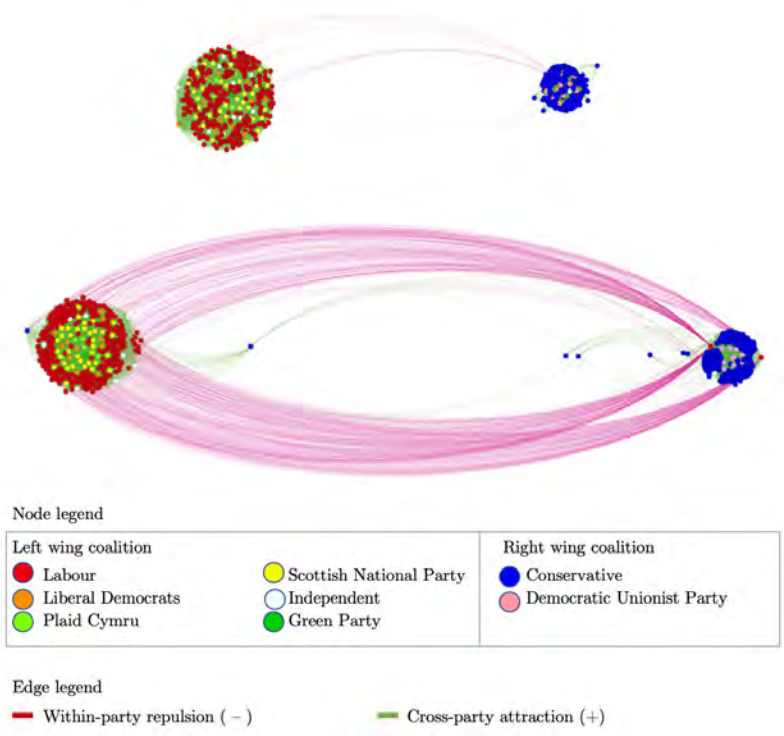


Fig. 1. Network projection of dissent and rebellion for non-Brexit (top) and Brexit (bottom) divisions, for the 8 political party case. Each node denotes an MP, connected by an edge to another MP. The colour of the edge represents whether the connection is a repulsive (pull) or attractive (push) force. Node colours denote party affiliation.

a cluster, it is evident that in the Brexit case, cross-party alliances could happen across two clusters that intuitively have polarised ideological beliefs.

We are able to focus on the results at the node level which allows us to investigate the MP's identified as rebels. We found that these MPs were the subsequent members of the breakaway party Change UK, a few notable MP "troublemakers" and some MPs who were faced with a moral decision to either support their party lines, or their people's vote.

In Figure 2(top), we replicate the visualisation of the two-ideology network, however with greater emphasis on the MPs that were identified as rebels.

Fig. 2. Upper panel: The rebels as identified by the network projection. The node colour represents ideological position: red for Left-wing; blue for Right-wing. Lower panel: The identified rebels using visual inspection, and rebellion score (highlighted in yellow).

References

1. Epstein, L.D.: Cohesion of British Parliamentary Parties. The American Political Science Review **50**(2) (1956) 360–377
2. Wood, D.M.: Comparing Parliamentary Voting on European Issues in France and Britain. Legislative Studies Quarterly **7**(1) (1982) 101–117
3. Moore, L.: Policy, Office and Votes: Conservative MPs and the Brexit Referendum. Parliamentary Affairs **71**(1) (January 2018) 1–27

Signed parliamentary networks: how frustration affects the government formation in parliamentary democracies *

Angela Fontan and Claudio Altafini

Division of Automatic Control, Department of Electrical Engineering,
Linköping University, SE-58183 Linköping, Sweden,
E-mail: {angela.fontan, claudio.altafini}@liu.se

1 Introduction

Multiparty parliamentary democracies are characterized by government formation processes that are often complex and protracted [1, 2]. Indeed a candidate cabinet needs the confidence of the parliament to rule, which implies that if there is no clear winner after the elections (i.e., no political party has secured a majority in the parliament) then a negotiation process has to start between the parties in order to achieve a coalition cabinet backed by the majority of the parliament.

In this work we consider the parliamentary elections in a total of 29 European countries, in a time interval that goes from 1978 to 2019. Differently from previous works [3, 4], we want to bring a new signed network perspective in the analysis of the government formation process, which we describe as a nonlinear model for decision-making [5–7] over signed networks: our aim is to predict both the duration of government coalition talks and the successful cabinet coalition outcome of the negotiations. We show that when the frustration (i.e., the amount of "disorder" [8]) encoded in the parliamentary network is high, then the period of negotiation talks between the parties will be long, with average values of correlation (duration of negotiation talks vs frustration) ranging from 0.42 to 0.68, depending on the amount of information encoded in the edges of the signed graphs. Moreover, the accuracy of our predictions on the cabinet composition after the elections ranges between 68.4% and 80.4%.

2 Methods

A parliamentary network is a graph where each node represents an elected Member of Parliament (MP). We assume that the network is complete, undirected and signed: each pair of MPs is linked by a (weighted) edge, whose sign represents the relationship between the two, cooperative (positive weight) or competitive (negative weight). If we gather all the MPs from the same political party in a single cluster node (assuming homogeneous party behavior), we obtain a clustered network $\mathscr{G}$, which translates into the adjacency matrix A of the network being a block matrix, as can be seen in Figure 1A. To construct the adjacency matrix we consider three scenarios, representing different

*Work supported in part by a grant from the Swedish Research Council (grant 2015-04390).

party grouping criteria and weight selection methods. **I**: unweighted "all-against-all" networks, that is, $a_{ij} \in \{-1,+1\}$ with $a_{ij} = +1$ only if the MPs i and j belong to the same party, see Fig. 1A. **II**: weighted according to the left-right position of each party in the political spectrum given by their electoral manifestos (the so-called "rile" index [9]), a positive weight is assigned between parties which formed a pre-electoral coalition. **III**: weighted according to a randomized and optimized left-right position, and pre-electoral coalitions. For each country and election, the signed parliamentary networks we obtain are in general not structurally balanced. Their frustration is calculated according to

$$\zeta(\mathscr{G}) = \frac{1}{2} \min_{\substack{S=\mathrm{diag}\{S_1,\ldots,S_{n_\mathrm{p}}\} \\ S_i=\pm 1\cdot I_{c_i}}} \sum_{i,j\neq i} [|\mathscr{L}| + S\mathscr{L}S]_{ij}, \tag{1}$$

where $|\cdot|$ indicates the absolute value, $\mathscr{L} = I - (\mathrm{diag}\{|A|\mathbb{1}\})^{-1}A$ is the normalized signed Laplacian of the network, n_p is the number of parties in the parliament, c_i is the number of seats won by the i-th party and S is a diagonal signature matrix ("spin" of the MPs). Our predicted government coalition, denoted $\mathscr{P}_{\mathrm{best,maj}}$, is given by the group of parties achieving a majority in the configuration S_best yielding the minimum in (1). Let $\mathscr{P}_\mathrm{gov}$ be the party coalition obtaining a confidence vote for the same election. The following index (here $\mathrm{card}(\cdot)$ indicates the cardinality of a set) evaluates the overlap between our prediction and the cabinet formed after the election

$$\rho_\mathrm{gov} = \frac{\mathrm{card}(\mathscr{P}_{\mathrm{best,maj}} \cap \mathscr{P}_\mathrm{gov})}{\mathrm{card}(\mathscr{P}_\mathrm{gov})}. \tag{2}$$

3 Results

For the 29 European nations of Fig. 1B, data for number of seats and position in the left-right political spectrum for each party, "rile" index, pre-electoral alliances, composition of the government formed after the elections and negotiation days were collected from various sources, such as WIKIPEDIA, the Manifesto Project Database [10], the Parliaments and Governments Database [11], and the new Parline (IPU's Open Data Platform) [12]. As we see in Fig. 1B the frustration correlates well with the duration of the government negotiation talks (calculated as the number of days from the general election to the date the government is sworn in) with average values ranging from 0.42 in scenario **I** to 0.68 in scenario **III**. Similarly, we obtain that our estimates represent well the actual cabinet composition, with average values for the index ρ_gov varying between 68.4% (scenario **I**) and 80.4% (scenario **III**).

References

1. S. N. Golder, "Bargaining Delays in the Government Formation Process," *Comparative Political Studies*, vol. 43, no. 1, pp. 3–32, jan 2010.
2. A. Ecker and T. M. Meyer, "The duration of government formation processes in Europe," *Research and Politics*, vol. 2, no. 4, pp. 1–9, 2015.
3. M. Debus, "Pre-electoral commitments and government formation," *Public Choice*, vol. 138, no. 1-2, pp. 45–64, jan 2009.

4. S. N. Golder, "Government Formation and Cabinets," in *Emerging Trends in the Social and Behavioral Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc., may 2015, pp. 1–15.
5. R. Gray, A. Franci, V. Srivastava, and N. E. Leonard, "Multiagent Decision-Making Dynamics Inspired by Honeybees," *IEEE TCNS*, vol. 5, no. 2, pp. 793–806, jun 2018.
6. A. Fontan and C. Altafini, "Multiequilibria Analysis for a Class of Collective Decision-Making Networked Systems," *IEEE TCNS*, vol. 5, no. 4, pp. 1931–1940, dec 2018.
7. ——, "Achieving a decision in antagonistic multi agent networks: frustration determines commitment strength," in *57th IEEE CDC*. Miami Beach, FL, USA: IEEE, dec 2018, pp. 109–114.
8. G. Facchetti, G. Iacono, and C. Altafini, "Computing global structural balance in large-scale signed social networks," *PNAS*, vol. 108, no. 52, pp. 20 953–20 958, dec 2011.
9. M. J. Laver and I. Budge, Eds., *Party Policy and Government Coalitions*. London: Palgrave Macmillan UK, 1992.
10. A. Volkens, W. Krause, P. Lehmann, T. Matthieß, N. Merz, S. Regel, and B. Weßels, "The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2019a." Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB), 2019. hhttps://doi.org/10.25522/manifesto.mpds.2019a.
11. D. Holger and P. Manow, "Parliaments and governments database (ParlGov): Information on parties, elections and cabinets in modern democracies. Development version." 2018. http://www.parlgov.org.
12. IPU, "New Parline: the IPU's Open Data Platform." https://data.ipu.org.

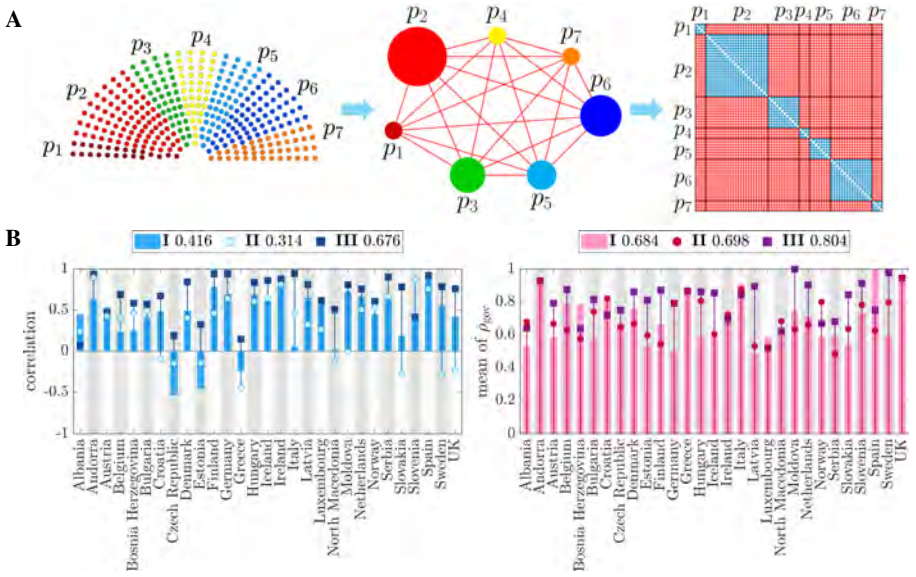


Fig. 1: (**A**): Constructing a parliamentary network (p_i = political party) and the corresponding adjacency matrix in scenario **I**: all parties against all parties and weights equal to $+1$ (blue) and -1 (red). (**B**): Results for the 29 countries and three scenarios (average values are reported above the plots). (left): Correlation between frustration and duration of government negotiation talks. (right): Index ρ_{gov}, i.e., overlap between predicted cabinet and party coalition in the government.

A complex network approach on the analysis of the Chilean presidential elections, using Twitter Data

Benjamín Ortiz E[1], Denisse Pastén[1], and Victor Muñoz[1]

Universidad de Chile, Las Palmeras 3425,
Ñuñoa, Santiago, Chile

1 Introduction

During election periods electoral analysis are common, usually via surveys, and maybe attempting to some kind of electoral forecast or prediction. In this work, we propose a new way to observe the political scene by using Twitter Data to build a complex network and using the weighted and non-weighted eigenvector centralities to measure the influence of each candidate before the election. This will not give us a prediction as such, but yield us important information about the analyzed system.
The proposed model consists of building a complex network from the interaction between Twitter users. The nodes of the complex network are Twitter users, and these are connected if they are mentioned or replied to each other. Also, the frequency of interaction between users is taken as the link weight.

We chose to analyze the Chilean presidential elections of 2017 which had the following key dates. August 21st: the electoral campaign period began, September 20th: the electoral propaganda period began; November 19th: first-round election, and December 17th: second-round election.
Using the scrapper GetOldTweets3 we downloaded Twitter data from August 21st of 2017 to December 20th of 2017, that is, from the start of the electoral campaign period, until a few days after the second round election. In particular were downloaded every tweet issued by the presidential candidates accounts, and every tweet that mentioned such accounts, or replied to them. Following we show the candidates names and their accounts: Eduardo Artés (@eduardo_artes), Marco Enriquez-Ominami (@marcoporchile), Carolina Goic (@carolinagoic), Alejandro Guillier (@guillier), José A. Kast (@joseantoniokast), Alejandro Navarro (@senadornavarro), Sebastián Piñera (@sebastianpinera), and Beatriz Sánchez (@labeasanchez).
In total, 605201 tweets were downloaded from 114761 different Twitter accounts.

Then, the Python module NetworkX was used to build the complex network during a given timeframe, calculate metrics for that timeframe, and follow the temporal evolution of such metrics. Specifically, we take time windows of 7 days length, moving in steps of one day. This in a similar way as in [3], but with larger networks that overlap each other. For each timeframe, we calculated the nodes degree, strength, eigenvector centrality, and the weighted eigenvector centrality.
We chose to calculate both eigenvector centralities as a measure of the importance or influence of each node (that is, of each Twitter user [4,5]) since we want to consider the indirect influence of nodes over the entire network, and not only over their first

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

neighbours. In particular we analyzed them for each candidate node.
The eigenvector centrality [1,2] can be calculated as

$$\mathbf{A} = \lambda \mathbf{x}$$

where $\mathbf{A}$ is the adjacency matrix, λ its greater eigenvalue, and $\mathbf{x}$ its respective eigenvector. Then, the centrality for the v^{th} node is the v^{th} entry of the eigenvector $\mathbf{x}$.
The Weighted version of the centrality is defined in a similar way, but using the weighted matrix instead.

2 Results

The following figures show the evolution of the eigenvector centrality and the weighted eigenvector centrality starting from the begin of electoral campaign period to the first round election day. We call this the first period.

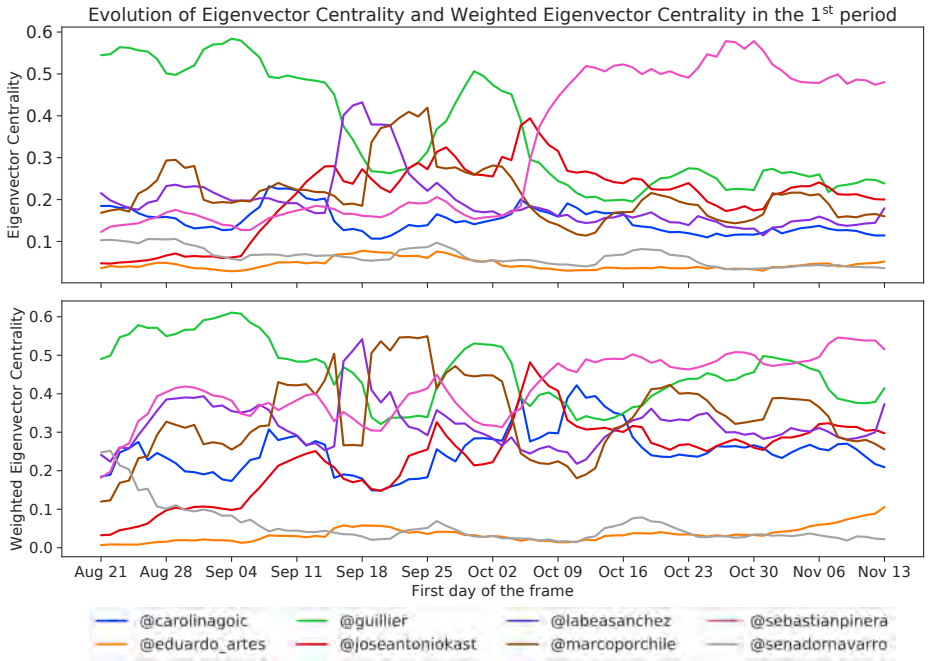


Fig. 1: Weighted and non-weighted wigenvector centrality measured in 7 days timeframes, moving with one day step, in the first period. The x axis shows the first day of the current frame.

We can observe that Twitter Data give us an interesting interpretation of the political scene, not too far from the reality. In fact, as the date approaches the election day, the order that the candidates centrality looks more like the order of the candidates in the election [6].

Then we consider the complex network in a second period, from the first round election day to December 20$^{\text{th}}$, getting the following figures.

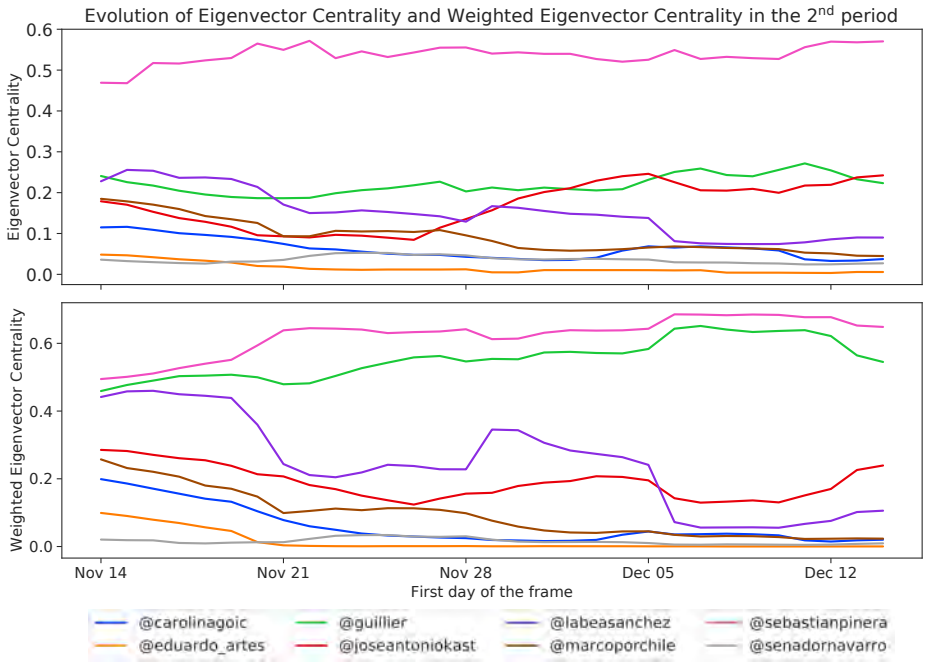


Fig. 2: Weighted and non-weighted eigenvector centrality measured in 7 days timeframes, moving with one day step, in the second period. The x axis shows the first day of the current frame.

In this period, the weighted centrality is more representative of reality because Alejandro Guillier, who also passed to second round election, shows a competitive score against Sebastian Piñera. Instead, in the non-weighted centrality he appears to not remain in competition. On the other hand, Sebastian Piñera near the first round election day, and then, in every moment previous to the second round election (Dec 11 frame), is the user with higher centrality, also he got the greatest amount of votes in both elections. Indeed, he finally got elected President of Chile.

References

1. Mark E. J. Newman: Networks: An Introduction. Oxford University Press, USA, 2010, pp. 169.
2. Bonacich, P. F., Power and centrality: A family of measures, Am. J. Sociol. 92, 1170-1182 (1987)
3. Braha, Dan and Bar-Yam, Yaneer. (2006). From Centrality to Temporary Fame: Dynamic Centrality in Complex Networks. Complexity. 12. 10.1002/cplx.20156.
4. Maharani, Warih et al. "Degree centrality and eigenvector centrality in twitter." 2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA) (2014): 1-5.
5. Dubois, E., and Gaffney, D. (2014). The Multiple Facets of Influence : Identifying Political Influentials and Opinion Leaders on Twitter. 1, 58, 1260-1277
6. 2017 Chilean Presidential Election Results. https://historico.servel.cl/servel/app/index.php?r=EleccionesGenerico&id=187

Coalitions and Coordination in Washington Think Tanks: Board interlock among Washington DC-based policy research and planning organizations

Alexander C. Furnas

University of Michigan, Ann Arbor MI, 48109, USA,
Email: zfurnas@umich.edu
Website: http://www.alexanderfurnas.com

1 Introduction

Governance requires massive, decentralized and coordinated information processing by governments. Lawmakers must evaluate the state of the world, identify issues worthy of government attention, prioritize among these issues, identify and evaluate plausible policy solutions, and then make a collective decision. In democratic systems, this must all be done while incorporating the preferences of lawmakers' constituents and other interest groups who both hold private policy-relevant information, and also have a stake in the outcomes of the governments decisions [11]. In the contemporary American context, lawmakers rely on a network of actors to gather and make sense of this diffuse information, and the output of these actors serves as inputs to the policymaking process.

As Bawn *et al.* articulate, we can understand political parties in the United States as extended networks of policy demanding constituencies, organizations and interest groups[1]. This theory is usually applied to the role these networks play in setting party agendas or nominating and supporting candidates for elected office. To be sure, the nomination and election process serves a fundamental role in identifying and prioritizing issues for government attention. However, I argue that policy research and planning organizations, often called "think tanks," serve the policy-apparatus of these extended party networks of policy-demanders. As such, we should expect them to coordinate along ideological dimensions in their attempts to support and influence partisan and ideological lawmakers[8]. While think tanks have been largely overlooked by scholars, recent work has begun to address their role in American Politics. What Think tanks pay attention to reflects partisan issue ownership[3]. Both think tank's ideological perspective and proximity to power shape how their work is used by Congress[6]. Moreover, congressional staffers both disproportionately trust policy evidence from partisan-aligned think tanks[4], and are more likely to favorable evaluate petitioners presenting evidence from aligned think tanks[5].

This paper is an inductive look at how these central—yet understudied—actors in the American political landscape coordinate among themselves. I conduct the largest mapping of the Washington D.C. think tank ecosystem to date[2]. Following an approach common in organizational sociology, I leverage interlocking directorates of organizations to examine patterns of organizational coordination[7]. Two organizations

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

are said to have an "interlock" between their directorates when one (or more) people sit on the boards of both organizations. This signifies a strong organizational tie between these two groups. I construct and analyze the board interlock network for 277 Washington D.C. based organizations, those with average annual budgets over $100,000 which are classified as subtype "Research Institutes & Public Policy Analysis" according to the IRS' National Taxonomy of Exempt Entities (NTEE) system. I obtain memberships of the board of directors for these organizations using from IRS Form 990 disclosures for 2008-2015 from GuideStar.

While I expect *a priori* that organizations will be more likely to foster ties with other organizations that are ideologically aligned with them and those that have access to greater financial resources, I do not have strong prior beliefs about the direction or magnitude of the relationships between other dyadic and organizational factors and the probability of connection. Common explanations for strategic interlock in the corporate context include preferences for diverse connections to increase information gathering breadth, and preferences for connections to other similar organizations to monitor competition, or engage in collusion[7]. Similar explanations map to this case: Organizations may seek to build cross-issue coalitions increasing the scope of their information networks (issue heterophily), or they may choose ties within their local issue space to increase efficiency and avoid duplicated effort (issue homophiliy). Similarly, we might imagine justifications for either homo- or heterophilic preferences in other dimensions of organization type, like whether the organization engages in lobbying, has dues paying members, or hires contractors.

2 Results

In this section, I present the results of an Exponential Random Graph Model on the think tank board interlock network shown in figure 1[9][10]. Edges are dichotomized to code for the existence of an interlock tie between organizations. Convergence and fit of this `ergm` model was good, but diagnostics are excluded from this abstract for space.

As expected, think tanks are substantially more likely to have interlocking boards as the revenue of the organizations increase; particularly successful fundraisers appear to be more attractive targets for interlock. Think tanks are less likely to be connected to those that are ideologically more distant from them. However, the magnitude of this effect is quite small (perhaps due to attenuation from measurement error induced by substantial missingness in my ideology measure). On the other hand, organizations are much more likely to interlock within their issue area (e.g. Medical Research; Agriculture, Food, Nutrition; or International, Foreign Affairs, and National Security). Think tanks are also more likely to interlock if they both engage in lobbying, and both engage in contracting. One possible explanation for this is that contractors and lobbyist may serve as potential vectors of informal relationships between organizations, which ultimately facilitates later institutionalization via interlock. However, organizations are substantially less likely to interlock if they both are membership organizations or are both non-membership organizations. Instead, organizations connect to those that are dissimilar along this dimension, suggesting that they find benefit in the differential expertise and resources that the other can provide. However, the amount of lobbying an

Think Tank Board Interlock Netv
Exponential Random Graph Mo

Parameter	Estimate	Stan
Ideological Distance	-0.0181	
Match NTEE	1.1681	
Match Membership Org	-0.3769	
Match Lobbying Org	0.0759	
Match Contracting Org	0.1564	
Log(Revenue)	0.2944	
Membership Dues	0.0000	
Lobbying Fees	0.0000	
Null Deviance	52992 on 3	
Residual Deviance	1782 on 3	

Table 1. Exponential Random Graph Model of D.C. Think Tank Board Interlock. Parameter estimates and standard errors rounded to 4 decimal places

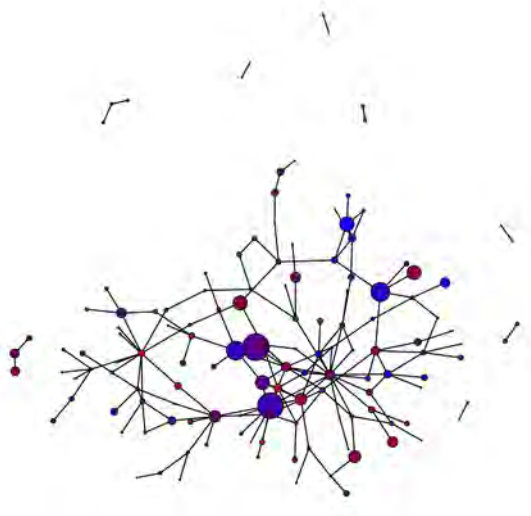


Fig. 1. D.C. Think Tank Board Interlock Network (isolates excluded)

organization does, or the membership dues it collects are not associated with whether these organizations connect. These results present a first, inductive look at coordination among understudied, elite policy planning organizations in the U.S.

References

1. Bawn, Kathleen, Martin Cohen, David Karol, Seth Masket, Hans Noel, and John Zaller. A theory of political parties: Groups, policy demands and nominations in American politics. Perspectives on Politics 10, no. 3 (2012): 571-597.
2. Burris, Val. The interlock structure of the policy-planning network and the right turn in US state policy. In Politics and Public Policy, pp. 3-42. Emerald Group Publishing Limited, 2008.
3. Fagan, E.J. 2019. Issue ownership and the priorities of party elites in the United States, 20042016. Party Politics. https://doi.org/10.1177/1354068819839212
4. Furnas, Alexander C. Biasing Their Bosses: Staff Ideology, Motivated Reasoning, and the Distortion of Information in Congress. Unpublished Working Paper. https://bit.ly/2ksg8i5
5. Furnas, Alexander C., Timothy LaPira, Alexander Hertel-Fernandez, Lee Drutman, and Kevin Kosar. Moneyed Interests, Information, and Action in Congress: A Survey Experiment. Unpublished Working Paper. https://bit.ly/2lwUYzE
6. Lerner, Joshua. Getting the Message Across: Evaluating Think Tank Influence in Congress. Unpublished Working Paper. http://sites.duke.edu/lerner/files/2015/04/Think_Tanks_PublicChoice_Revised.docx
7. Mizruchi, Mark S. What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. Annual review of sociology 22, no. 1 (1996): 271-298.
8. Noel, Hans. Political ideologies and political parties in America. Cambridge University Press, 2014.
9. Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. Social networks 29, no. 2 (2007): 173-191.

10. Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. Journal of statistical software 24, no. 1 (2008): 1548.
11. Hayek, Friedrich August. The use of knowledge in society. The American economic review 35, no. 4 (1945): 519-530.

Reply networks on Twitter

Felix Gaisbauer, Armin Pournaki, Sven Banisch, and Eckehard Olbrich

Max Planck Institute for Mathematics in the Sciences, Leipzig
felix.gaisbauer@mis.mpg.de

1 Introduction

User interaction on social media platforms has been used extensively to monitor and analyse political discourse and the spectra of political opinion, serving as an empirical basis for modelling and investigating opinion dynamics. Due to the accessibility of data and the 'public-by-default' nature of the platform, Twitter serves as the first choice for many researchers. But most previous studies on polarization and user interaction on Twitter have mainly focused on the retweet network of posts concerning a certain topic or coming from a specific base set of users. There, separate clusters of users can very often be identified, and users of a cluster might then be assigned a certain political leaning on the topic under investigation.[1, 2] Moreover, very weakly connected clusters are often taken as evidence for a very polarized political discourse.

While this procedure might produce meaningful results in terms of identifying political coalitions on social media (retweets are seen as endorsements of opinions or appreciation and distribution of information/arguments), it ignores direct interaction, that is, replies between users, which can be positive or negative. We will show in this contribution that a retweet network alone often lacks crucial information about political discourse, and that a reply network, constructed from the replies of users between each other, provides valuable insight that should not be neglected. We analyse tweets about the Saxonian state elections of 2019 in order to substantiate this claim. We find that, contrary to the clustering found in the retweet network, direct user interaction in the form of replies is strong between the different retweet clusters; for the nature of this interaction, a classification of user opinion based on the retweet network can serve as a proxy: Different-cluster interaction might in most cases be characterized by disagreement or negative feedback, while same-cluster interaction signals agreement. Interestingly, the direct feedback is asymmetric in the sense that many users of one cluster (associated to the AfD party) tend to reply to users of the other, while the reverse is not true. We conclude that, by only considering retweet networks, one ignores a mode of interaction which is not related to argument or information diffusion, but rather to accessing and controlling public political discussion.

2 Results

We collected Twitter data about the Saxonian state elections which took place on September 1, 2019. We used a seed sample of 208 Twitter users composed of candidates of the most prominent political parties taking part in the election, bloggers and journalists with

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

a focus on Saxonian politics, and users that contributed to discussions on the election extensively on Twitter. We collected, in the three months preceding the elections, all tweets produced by this group as well as all tweets that mentioned said users in any way: Retweets, replies and mentions. We constructed a retweet network out of all collected retweets. Using Force Atlas 2, a physics-based layout algorithm,[3] we arranged the network in a two-dimensional space, arriving at two separate clusters; we classified the users of the two clusters accordingly. In the first one (blue in Figure 1), mainly politicians of the AfD party and users retweeting their content were placed; in the other big cluster, politicians of the other big parties were found (red).[1] We then built all the complete reply trees in the sample of the last month preceding the election and visualized them, as shown in Figure 1. The cluster classification now served as a proxy for the type of interactions that took place in a reply tree: A user of one cluster replying to a user of another cluster probably disagreed with the opinion of the latter, while chains of posts from users of the same cluster were interpreted as agreement. Around 65 percent of the tweets and 50 percent of the users involved in the reply trees could be classified.

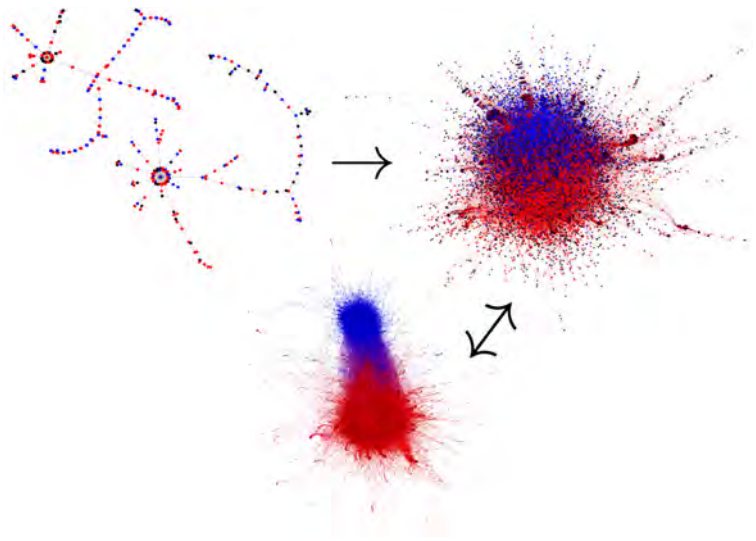


Fig. 1. Exemplary reply trees (top left) and the corresponding reply network (top right). The corresponding retweet network, upon which the classification indicated by the blue and red colors was performed, is shown below. The black vertices in the reply network are users which did not show up in the retweet network and hence could not be classified.

Previous research on reply trees has mainly focused on generative models (see [4] for a review). The above classification could enrich generative models as a feature, but also might substitute as a proxy, in suitable cases, computationally and effort-costly

[1]Note that this is a tentative clustering procedure that we seek to systematize soon, giving conditions under which this interpretation is valid. Alternative community detection measures such as modularity separate the two clusters as well, but in more than two communities.

approaches to gain insight into the type of user interaction such as sentiment analysis. Moreover, a reply network can be constructed out of the reply data, assigning a directed edge between two users if one has replied to the other. The structure of this network is fundamentally different from the retweet network: We have many cross-group interactions between users. This yields (again using Force Atlas 2) a non-polarized graph (cf. Figure 1), in which the users are arranged in one big cluster. The different 'opinion clusters' are asymmetric in their interaction patterns: While the users of one cluster reply to users of the other group very frequently (much more often than among each other), it is the opposite for the other cluster (see Table 1).

Table 1. The number of users and replies within and between the different opinion clusters.

	# of users	# of replies	# of repl. within cluster	# of repl. to other cluster
Cluster 1	3334	16465	4375	9153
Cluster 2	5001	25085	13468	4039

Hence, retweeting happens in relatively isolated clusters. But direct interaction, that is, discussion among users, does not. The relation between the two calls for further interpretation. One hypothesis could be that argument and information diffusion occurs within the own opinion community, while the competition for public opinion dominance really is *public*. For modelling approaches in opinion dynamics, this should be taken into account and might point into the direction of multi-layered interaction. Moreover, this contribution serves as an example that different networks can enrich each other; here, the retweet network presents only a one-sided view on online communication, while the reply network is particularly insightful if assisted by information from the structure of the retweet network.

Summary. We analysed the retweet as well as the direct reply interactions between users on Twitter for the 2019 Saxonian state elections. We found that while the retweet network analysis indicates polarized opinion clusters, user interaction in terms of direct replies does not. There is interaction between users placed in separate retweet clusters, which, interestingly, is asymmetric in the sense that one cluster replies much more often to the other than to users of the own; the reverse is not the case. These results need interpretation and might influence future modelling efforts.

References

1. Conover et al.: Political Polarization on Twitter. Fifth International AAAI Conference on Weblogs and Social Media (2011)
2. Wong et al.: Quantifying Political Leaning from Tweets, Retweets, and Retweeters. IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 8 (2016)
3. Jacomy et al.: ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9(6): e98679 (2014)
4. Aragon et al.: Generative models of online discussion threads: state of the art and research challenges. Journal of Internet Services and Applications, Vol. 8 (2017)

Polarisation and complexity in parliamentary debates

Paulo Almeida[1], Lília Perfeito[1] Manuel Marques-Pita[2], and Joana Gonçalves-Sá[1]

[1] Nova School of Business and Economics, Carcavelos, Portugal,
joana.sa@novasbe.pt,
http://scienceandpolicy.eu/
[2] Universidade Lusófona, Lisboa, Portugal

1 Introduction

Recently, much attention has been given to what is perceived as a strong increase in political polarisation (between parties) and decreased heterogeneity (within parties). Such diversion is argued to happen both at the discourse and voting levels. However, this has mostly been studied through quantitative analysis of voting patterns, particularly in two-party systems, such as the one found in the USA, where representatives have independence of vote, with the underlying assumption that **discourse and voting patterns are aligned**. As quantitatively studying discourse or subtle forms of political dissent is difficult; most analysis have been limited to small scale case studies or voting networks of two-party systems. However, the majority of parliaments have more than two parties and several have limited freedom of vote, hence there are clear limits to the generalisation of current tools, mostly developed to study the US system.

Here, we focused on the Portuguese Parliament (PP) as it offers interesting challenges both at the political science and computational levels, representing an excellent case study to approach multi-party systems. From its peaceful revolution in 1974, to joining the EU, Portugal has had some very rich 40 years of political history, while maintaining a democratic regime. It is a relatively small parliament, with a varying number of parties, that does not favour voting freedom. The two major parties (centre-left PS and centre-right PSD), both considered moderate, alternate in forming government, with the support of (or in coalition with) smaller parties to the right (CDS) or to the left (BE and PCP).

Here, we describe a novel quantitative and computational approach to the study of parliaments and to test the assumption of alignment between voting and discourse. First, we extracted all plenary debates of the PP for the past 40 years, and built a pipeline to automatically identify all speakers and parties. To do this, we created a network of MPs, including information such as gender, age, party, or formal education; we curated all plenary debates, and created a pipeline for automatically identifying the speakers on all public sessions. All discourse for each speaker (when possible) and party, was then analysed using readability [1]) as a complexity measure.

In parallel, we scraped the Parliament website to get votes for all available legislative initiatives. These votes are tallied by party, with only the MPs who diverge from their party being named. We analysed voting patterns using Multiple Correspondence Analysis (MCA) [2], which allowed us to reduce dimensionality and measure changes in distance between parties over time.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

We find that different parties can have almost completely aligned political positions (as gauged from voting patterns) while having substantial differences in discourse. Together, and to our knowledge, our work provides the first example of a large-scale comparison between speech and votes, and a new a tool that can be applied to the study of different multi-party systems, offering important insight into the differences between discourse and actual policy

2 Results

We started by investigating whether we could measure an increase in polarisation, or voting distance, with time. Two lines of evidence point to such an increase. First, the proportion of bills that passed without any vote against decreased from 85 % in the early 1980s to 31% in the latest legislature (binomial regression $p < 0.0001$). Next, we compared the co-voting frequency of all parties with the Communist party (PCP). We chose PCP because they have had parliamentary representation since the first legislature, and because they have traditionally always been among the most left-leaning MPs. As can be seen in figure 1A, the frequency of co-vote between PCP and the two traditionally right-leaning parties (CDS-PP and PSD) has been steadily decreasing (logistic regression $p < 0.0001$ for either PSD and CDS-PP).

In order to go beyond parwise comparisons, we used MCA to reduce the dimensionality of voting patterns. We used the MCA python implementation developed by [3]. Each party is classified as either voting for, against or abstaining for each bill presented in each legislature. Bills where not all parties were recorded were removed from the analysis (about 20 % over all legislatures). The MCA algorithm then extracts the factors that capture most of the variation in the data (similarly to principal component analysis but using categorical data). For all legislatures the first factor captured at least 70% of the variance and this is the factor we show in figure 1B.

Aside from the increased polarity in voting, the analysis also revealed that the two leftmost parties, PCP and Bloco de Esquerda (BE) almost always vote together (on average 85 %). This raises the question of how these two parties are both kept in parliament when they make the same decisions regarding bills. There are at least two hypotheses to explain their maintenance. One is that they appeal to different constituencies by discussing policies differently. The other is that the few bills where they disagree are enough to polarise the left-wing voters. To test the discourse hypothesis, we looked at speech complexity, measured as readability. A text with high readability has few words per sentence and few syllables per word. Originally this measure was devised for English [1] as a linear scale from 0 to 100 where texts higher than 50 are considered to be readable by people who finished high school and texts below 30 only by college graduates [4]. We validated this measure for Portuguese (not shown) and tested it in the Portuguese Parliament discourse using the adaptation proposed by [5]. Figure 1C shows readability over time for all parties. In general, readability is low in the PP (between 30 and 50). Surprisingly, readability for PCP is quite low, despite the fact that Communist MPs on average attended fewer years in University, and much higher for the other left-leaning party, BE. This suggests that readability is perhaps indicating more about the MPs' speech proficiency and less about who their target audience is. In either case,

readability indicates that PCP and BE are probably being understood differently by voters. In general, we can see in figure 1D that speech and votes are not always aligned. Indeed, data shows that the parties with the higher readability are parties which are close to either an older party (PCP) or bigger party (PSD) and hence may need to differentiate themselves to appeal to their electorates.

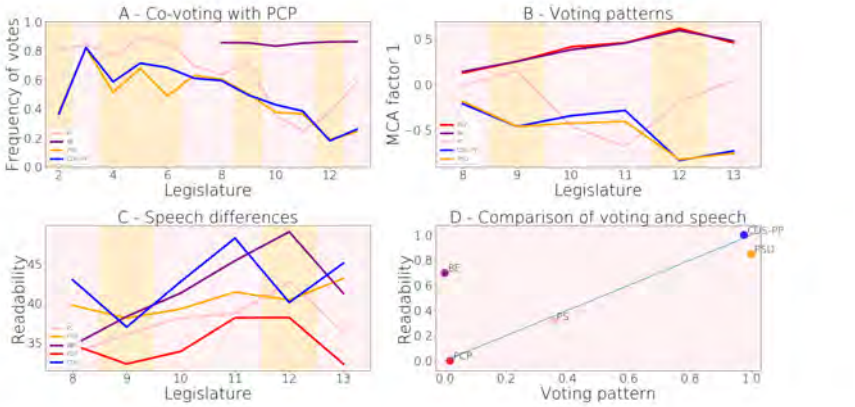


Fig. 1. Comparison of speech and voting patterns in the Portuguese Parliament. In all cases the color of the background represents which of the two major parties was in government, either center-left PS (pink) or center-right PSD (orange). A: Frequency of co-votes with PCP per legislature, per party. B: First factor in MCA per party per legislature. C: Readability per party per legislature. D: Correlation between voting patterns (measured by MCA factor 1) and readability in the latest legislature. To facilitate the comparison, both axes were normalised.

We demonstrate that, in the Portuguese Parliament, divergence between left and right is increasing. In contrast, two left-wing parties with almost identical voting patterns have been coexisting for 20 years. By measuring speech complexity, we suggest that these two parties are appealing to different voters. In general, we propose a methodology to compare voting and speech and show that parliamentary debates can be used to understand speech dynamics within a human network.

References

1. Flesch, R. A new readability yardstick. Journal of applied psychology, 32(3), 221 (1948)
2. Michael Greenacre, Jrg Blasius: Multiple Correspondence Analysis and Related Methods, CRC Press. (2006) ISBN 1584886285.
3. https://github.com/esafak/mca
4. Flesch, Rudolf. "How to Write Plain English". University of Canterbury. (2016)
5. Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. das G. V., & Oliveira Junior, O. N. de. Readability formulas applied to textbooks in brazilian portuguese. Sao Carlos: Icmsc-Usp. (1996)

Evolution of alliance and rivalry networks in international relations

Koji Oishi[1] and Kentaro Sakuwa[2]

[1] Global Research Center for Big Data Mathematics, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
oishi@nii.ac.jp

[2] Department of International Politics, Aoyama Gakuin University, 4-4-25 Shibuya, Shibuya-ku Tokyo 150-8366, Japan
sakuwa@sipeb.aoyama.ac.jp

1 Introduction

International relations are a typical example of signed networks that evolve over time as some countries are friends while others are foes and relations can quickly turn from cooperative to antagonistic and vice versa. Underlying mechanisms of the dynamics have not been fully understood while it often has a significant impact on international security. A well-known theory for the evolution of singed social networks is Heider's structural balance theory, which was originally proposed in social psychology and later translated into and studied as mathematical models of networks (e.g., [1]). According to the balance theory, signed social networks evolve to increase balanced triads, in which either the enemy of your enemy or the friend of your friend is your friend. In other words, triads with zero or two negative edges (hereafter denoted by $+++$ and $+--$ triads, respectively) are balanced, while those with one or three negative edges ($++-$ and $---$, respectively) are imbalanced, and balanced triads are expected more stable than imbalanced ones.

The balance theory is supported in various studies of social networks [2, 3]. However, few studies have tested the balance theory in international relations [4]. Therefore, we investigated the network of international alliances and rivalries and examined whether its evolution from 1816 to 2009 is consistent with the balance theory. We constructed the network of alliances and rivalries between sovereign states for each year between 1816 and 2009, by combining datasets about the membership of the sovereign state system [5], alliances [6], and rivalries [7]. Nodes are the sovereign states that existed in the year. Two sovereign states have a negative edge if they have a rivalry in the year while they have a positive edge if they do not have a rivalry and have a military alliance (either defensive or offensive) in the year.

2 Results

We found the evolution of the alliance and rivalry network is clearly different across three periods, 1816–1866, 1867–1941, and 1942–2009 (Fig. 1). While the number of nodes (sovereign states) increased over all the periods, the average degree dropped

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

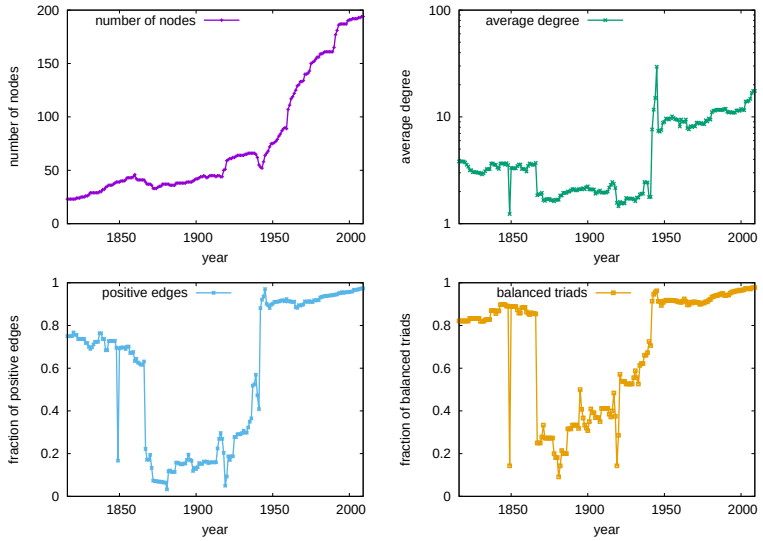


Fig. 1. The number of nodes (left top), the average degree (right top), the fraction of positive edges (left bottom), and the fraction of balanced triads (right bottom).

down in 1867 and jumped up in 1942. By the same token, the fraction of positive edges and the fraction of balanced triads ($+++$ and $+--$) are constantly high (i.e., positive edges and balanced triads are the majority) during 1816–1866 and 1942–2009, while both fractions suddenly dropped down in 1867 and gradually increased between 1867–1941. Therefore, the balanced triads are not always the majority and their fraction did not always increase.

Furthermore, we compared the empirical network with surrogate networks in which signs of edges are randomly shuffled without changing the topology of the network (Fig. 2). During the first (1816–1866) and third (1942–2009) periods, the empirical network and the surrogate networks are clearly different and the difference is consistent with the balance theory (i.e., balanced triads are more common than expected from the network topology and the fraction of edge signs). On the other hand, the difference between the empirical and the surrogate in the second period is much less clear, therefore the support for the balance theory is weaker in the period.

Our analysis revealed that the consistency of the balance theory with the empirical evolution of the international network of alliances and rivalries totally depends on the period. A possible reason is that sovereign states can split and merge, while previous studies that support the balance theory examined social networks of individuals [2, 3], in which split and merger of nodes are unlikely. For example, when nodes densely connected by positive edges merged into a new node, the balance of the network can drastically change. It is also likely that war plays a key role as the beginning of the third period (1942–2009) is during the Second World War. This study implies that we need to incorporate not only the balance theory but also these additional factors when we model the long-time evolution of international relations.

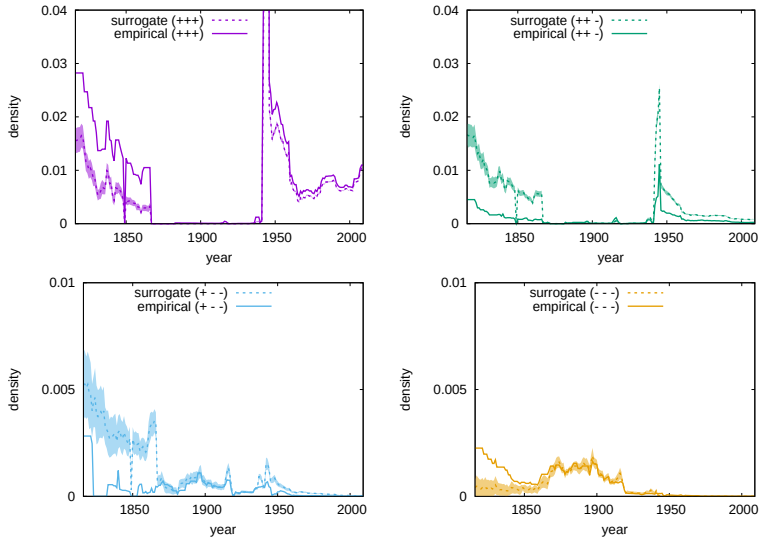


Fig. 2. The density of the triad of each type ($+++$, $++-$, $+--$, and $---$). The density is the ratio of the number of the triads of the type to the number of all possible combinations of three nodes, i.e., $N(N-1)(N-2)/6$, where N is the number of nodes. Solid lines represent the values for empirical networks and broken lines represent those for randomized networks averaged over 100 samples. Shaded areas show the standard deviation.

Summary. We investigated the evolution of the signed network of alliances and rivalries in international relations. The balanced triads are dominant and clearly more common than randomly signed networks in 1816–1866 and 1942–2009, consistently with the balance theory. On the other hand, the difference with randomly signed networks is not clear and the imbalanced triads are dominant in 1867–1941. The result shows the balance theory is only partly supported, implying that we need to incorporate other factors when we build a generative model of the evolution of international relations.

References

1. Antal, T., Krapivsky, P. L., Redner S.: Dynamics of social balance on networks. Phys. Rev. E 72, 036121 (2005)
2. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1361–1370 (2010)
3. Szell, M., Lambiotte, R., Thurner, S.: Multirelational organization of large-scale social networks in an online world. Proc. Natl. Acad. Sci. USA 107(31), 13636–13641 (2010)
4. Maoz, Z. *et al.*: What is the enemy of my enemy? Causes and consequences of imbalanced international relations, 18162001. The Journal of Politics 69(1), 100–115 (2007)
5. Correlates of War Project. State System Membership List, v2016. (2017)
6. Leeds, B., *et al.*: Alliance Treaty Obligations and Provisions, 1815-1944. International Interactions 28(3), 237–260 (2002)
7. Dreyer D., Thompson, W. R.: Handbook of International Rivalries. CQ Press (2011).

Dynamics of Commenters' Networks across Time and Political Spectrum

N. Gizem Bacaksizlar[1] and Mirta Galesic[1]

Santa Fe Institute, Santa Fe, New Mexico, 87501, USA
gizem@santafe.edu, galesic@santafe.edu

1 Introduction

Millions of people comment daily on current events on a variety of platforms ranging from diverse social media to the news sites themselves. Reading others' comments can shape one's own opinions about the story, its author, and the media outlet (see [4]), and help spread opinions and claims which counter the mainstream narrative [5]. Here[1] we investigate the social network of commenters on four well-known U.S. news sites that span the political spectrum from left to right. Studying the structure of the commenters' social network can reveal whether discussions are influenced by relatively few commenters, or whether they are shaped by many different commenters and thus more in line with Habermasian ideal of public discourse [1].

We focus on two factors that could affect the dynamics of influence in commenters' social networks. The first factor is the political orientation of a site's audience. Right-leaning political orientation has been linked to increased deference to authorities [3]. Consequently, social networks of commenters on right-leaning sites might be characterized by a larger inequality between commenters in terms of their influence on others, with some being disproportionately influential.

The second factor is the contemporaneous societal situation that can pose more or less threat to a certain side of the political spectrum. For example, just before the 2016 U.S. Presidential Election, Clinton was favored as the election winner, which might have been perceived as a threat among the Trump-supporting voters. In contrast, after the 2016 election and Trump's victory, the Clinton-supporting voters might have felt threatened. Studies show that groups under threat tend to become more homogeneous and follow thought leaders [2, 6], suggesting that some commenters might become much more influential than others as the perceived threat to their group increases.

2 Methods

We collected all comments posted a month before U.S. 2016 Presidential Election from October 7 to November 8 and a month after from November 8 to December 9 to four U.S. news sites, including two on the left side of the political spectrum - Mother Jones

[1]This work was supported by a grant from the National Science Foundation (DRMS 1757211). The funder had no role in study design or interpretation of results. We thank Joshua Garland, Kenan Turbic, and Henrik Olsson for helpful discussions.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

(M) and Atlantic (A), one more moderate - Hill (H), and one on the right side - Breitbart (B). All sites used Disqus commenting platform. We analyzed all comments posted to an article on the day the article first appeared. Table 1 shows the number of unique commenters and comments included in the analysis.

Table 1. Commenter and Comment Counts from Different News Sites and Time Periods

	Before Election		After Election	
News Site	**Commenters**	**Comments**	**Commenters**	**Comments**
Mother Jones (M)	4,504	75,240	4,839	70,852
Atlantic (A)	10,512	170,048	9,545	138,960
Hill (H)	33,055	1,250,003	34,861	1,054,394
Breitbart (B)	50,263	1,747,853	59,579	1,857,076

The social network of commenters is directed, with edges occurring when one commenter responds to the comment of another. Edges are weighted by the number of times each two pairs of commenters replied to each other in a given time frame. We investigate two measures of commenters' influence. The first is commenters' in-degree, the number of unique commenters who replied to their comments within a given time period. We normalize the distribution of in-degrees d by the number of commenters N on a site s in time period t $(d_{s,t}/(N_{s,t}-1))$. To compare proportions of commenters with different degrees $p(d)$ on different sites, we normalize them by the lowest proportion of commenters with a certain degree on each site $(p(d_{s,t})/min(p(d_{s,t}))$. The second measure is commenters' PageRank centrality PR weighted by degree weights, representing a more nuanced measure of commenters' importance than in-degrees. To compare PRs across sites with different number of commenters, we normalize them with the lowest possible PR for a given site and time period $(PR_{s,t}/min(PR_{s,t})$.

3 Results

Is inequality of influence larger on right-leaning sites? Our results suggest a more nuanced pattern. In-degree analysis (Fig 1 left) shows somewhat larger in-degrees for left-leaning sites (M and A) compared to moderate (H) and right-leaning (B) sites. Inequality of in-degree distributions is larger on the most left-leaning site (M) than on the most right-leaning site (B) (Fig 2 left). PageRank analysis (Fig 1 right) suggests that moderate and right-leaning sites have more very influential commenters. Furthermore, inequality of influence is higher on both extremes of the political spectrum (M and B) compared to more moderate sites (Fig 2 right).

Is inequality of influence larger when a group feels threatened? Our results offer some support for this hypothesis. Overall, commenters on right-leaning B have reliably larger in-degrees on average before than after the election (KS test, $p < .001$, Fig 1 left). In addition, analysis of distributions of in-degrees and PageRank indices for top 100 commenters (Fig 2) suggest that inequality of these distributions tends to increase for left-leaning sites and decrease for moderate and right-leaning sites from before to after the election.

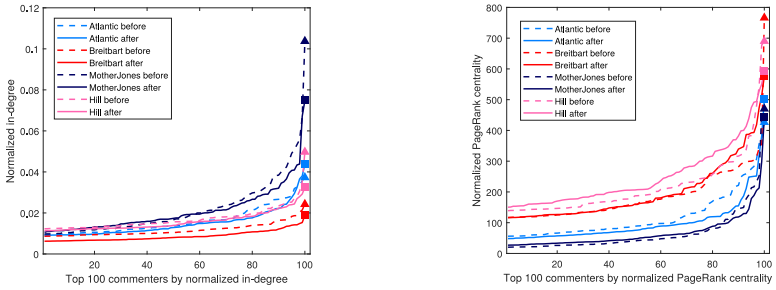


Fig. 1. Distribution of Influence for top 100 commenters: Normalized in-degrees (left) and PageRank centralities (right), before (triangles) and after (squares) the election.

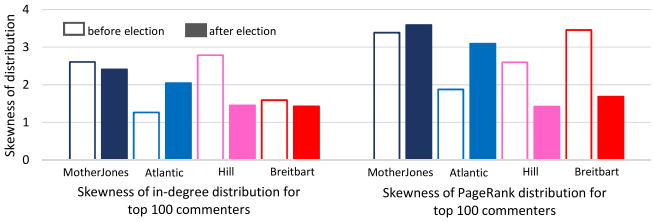


Fig. 2. Inequality (measured as skewness) of in-degree (left) and PageRank distributions (right) for top 100 commenters, before (empty bars) and after (full bars) the election.

4 Discussion

Our results suggest that commenters on political sites on the extremes of political spectrum tend to have a higher inequality of influence than commenters on more moderate sites, with some commenters being much more influential than others. Furthermore, we find the tendency for higher inequality of influence at times when supporters of a particular site feel threatened. Further work will explore the characteristics of most influential users and the role of commenters posting normative or antagonistic comments.

References

1. Habermas, J.: The structural transformation of the public sphere: An inquiry into a category of bourgeois society. MIT press (1991)
2. Janis, I.L.: Groupthink: Psychological studies of policy decisions and fiascoes, vol. 349. Houghton Mifflin Boston (1982)
3. Jost, J.T., Glaser, J., Kruglanski, A.W., Sulloway, F.J.: Political conservatism as motivated social cognition. Psychological Bulletin 129(3), 339 (2003)
4. Prochazka, F., Weber, P., Schweiger, W.: Effects of civility and reasoning in user comments on perceived journalistic quality. Journalism Studies 19(1), 62–78 (2018)
5. Toepfl, F., Piwoni, E.: Public spheres in interaction: Comment sections of news websites as counterpublic spaces. Journal of Communication 65(3), 465–488 (2015)
6. Turner, M.E., Pratkanis, A.R., Probasco, P., Leve, C.: Threat, cohesion, and group effectiveness. Journal of Personality and Social Psychology 63(5), 781 (1992)

What is going on Brazil? A Political Tale from Tweets

Diogo Pacheco, Alessandro Flammini, and Filippo Menczer

Center for Complex Networks and Systems Research
Indiana University, Bloomington, IN, USA

1 Introduction

Brexit in Europe, Trump in the U.S., and Bolsonaro in Brazil are examples of the increasing polarization of political debates across the world [2]. Concurrently, we have been witnessing the key role played by online social platforms as they become the main media for campaigns, debates, and recruitment [1, 4, 5].

Here we use social media data and network analysis to understand and highlight population-level political behavior in Brazil, as groups evolve from campaign competitors to new government and opposition blocks. Our analysis reveals a transition from a phase before the vote with many polarized groups to a phase after the vote in which these votes coalesced around a government and an opposition cluster.

2 Results

We used the Twitter streaming API to track the 14 Brazilian presidential candidates of the 2018 elections and the Brazilian Superior Electoral Court (`@TSEjusbr`). For each candidate, we followed four terms: the official account ID and handle (e.g., `@jairbolsonaro`), a hashtag associated with the campaign (e.g., `#Bolsonaro17`), and the full name (e.g., "Jair Messias Bolsonaro"). This yielded a collection of 104 million tweets from 3.8 million accounts between Aug. 30, 2018 and Aug. 26, 2019.

The Twitter timeline in Fig. 1A highlights some visible changes in political engagement. The activity increases until election day, followed by a drop in the period between the election results and inauguration day, and finally, the number of tweets and users stabilizes with few peaks around important events.

Fig. 1B shows a 10% drop in the retweet rate from 71% before the final election to 61% after inauguration day. The inset shows a relative drop of 60% in the number of original tweets. The number of replies, on the other hand, more than doubled, from 14% to 30%. These changes may represent two distinct behaviors: propaganda during the campaign, and debate during the mandate.

Despite the steady daily activity, Fig. 1C shows that nearly five thousand new accounts join the Brazilian political conversation every day. This suggests an account churn rate of roughly 5%. In future research, we plan to investigate who are the accounts leaving the conversation. One possible interpretation is that the dynamics are driven by many bots, which are replaced by new ones when they are suspended by the platform. In the present analysis we did not evaluate bot activity.

To analyze polarization, we created daily weighted networks in which nodes represent accounts and an edge connects two accounts if one mentions, retweets, quotes, or

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

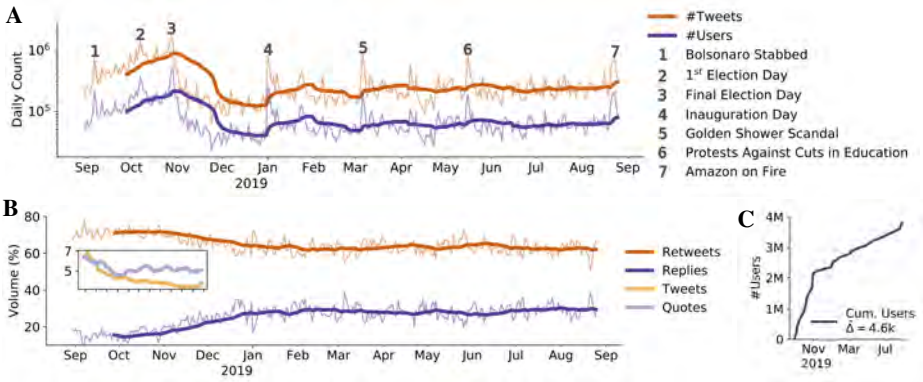


Fig. 1: **A Political Tale from Tweets.** (**A**) One-year Twitter timeline related to Brazilian politics. Peaking days are highlighted, e.g., Bolsonaro's murder attempt (**1**) and the worldwide protests against Amazon forest fires (**7**). The number of tweets and unique users is correlated ($\rho = .95$). (**B**) 30-day moving average of percentages of types of messages. *Retweets* and *replies* are negatively correlated ($\rho = -.94$). (**C**) Growth in number of users involved in the Brazilian politics conversation. The median growth is 4.6k new users per day.

replies to the other; the weight represents the number of interactions in a day. We also considered longer periods, with similar results.

Fig. 2 displays examples of the $k = 15$ network core, with colors representing high-modularity communities. Fig. 2A shows that on the eve of the first-round election modularity was high, with several clusters representing the candidates and their supporters. The inauguration day network (Fig. 2B) has lower modularity, but we can identify clusters corresponding to the new government (blue), the opposition (green), and the international community (pink and orange) bridged by the Brazilian Minister of Foreign Affairs (`@ErnestoAraujo`). The last example shows the network on August 23, 2019, during worldwide protests against Brazilian agro-business policies and the fires in the Amazon forest. The French president was heavily criticized by supporters of the Brazilian government for posting a photo of the fires that was later revealed to be old and not from the Amazon. That is why `@EmmanuelMacron` appears in the center of the network.

Fig. 2D shows that network modularity was highest before the election, during the campaign. A similar pattern has been observed in data about Italian elections [3]. After fluctuations over the first quarter of the new government, the modularity seems to be increasing again. This could be an indicator of growing government disapproval or an early sign of the next campaign cycle.

Finally, we analyzed how the Brazilian political scenario attracts international attention. We measured the percentage of messages per country and per day, using the country code metadata present in some of the tweets. Fig. 2E shows the standardized attention timeline, computed by the z-score of the relative volumes, for USA, Chile, Argentina, France, and Portugal. The timelines highlight the days with most activity for each country, suggesting that international attention was unevenly distributed due to distinct events. For instance, attention from the U.S. peaked on inauguration day, whereas attention from France peaked around the Amazon fires.

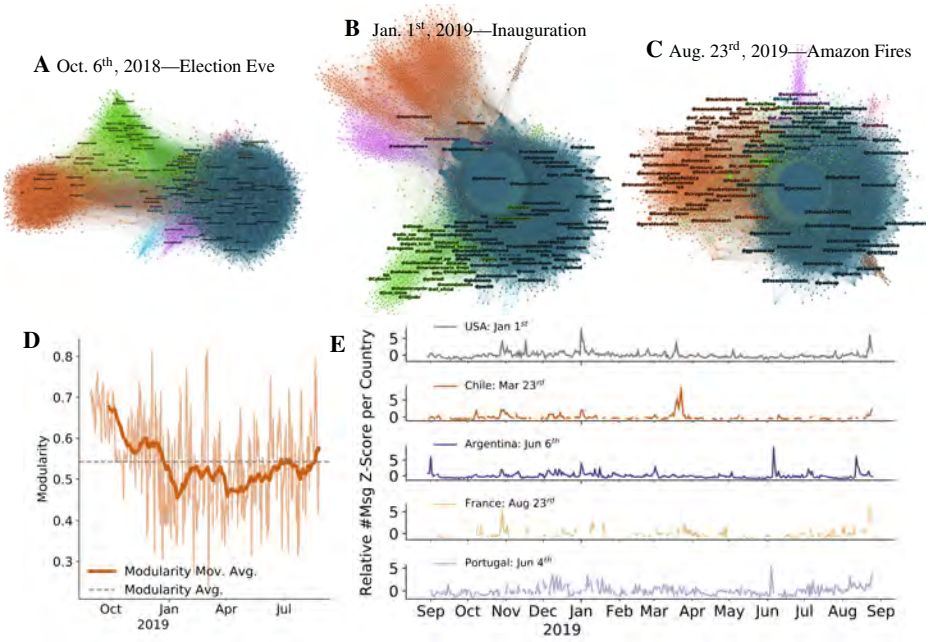


Fig. 2: **Network Perspective.** Examples of $k = 15$ discussion network cores: (**A**) The eve of election's first-round; (**B**) Inauguration day; (**C**) Amazon fires. (**D**) 30-day moving average of daily network modularity. (**E**) Peaks in international attention towards Brazilian politics.

Summary Our analysis reveals two distinct phases in the online discussion around the Brazilian election: a first phase prior to the vote, dominated by retweets and with higher degree of polarization around the many candidates; and a more conversational phase following the vote, with two main clusters around the government and the opposition. Although this analysis is based on a huge sample of online users, we do not know how representative Twitter data is of the Brazilian political spectrum.

References

1. A. Badawy, E. Ferrara, and K. Lerman. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265, 2018.
2. M. A. Eger and S. Valdez. Neo-nationalism in Western Europe. *European Sociological Review*, 31(1):115–130, 2015.
3. M. Lai, V. Patti, G. Ruffo, and P. Rosso. Stance evolution and Twitter interactions in an Italian political debate. In *Lecture Notes in Computer Science*, volume 10859, pages 15–27. Springer, 2018.
4. S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
5. K. C. Yang, P. M. Hui, and F. Menczer. Bot Electioneering Volume: Visualizing social bot activity during elections. In *The Web Conference Companion*, pages 214–217, 2019.

Part XVI

Quantifying Success

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Quantifying predictability in Football through network analysis; A historical approach

Victor Martins Maimone[1] and Taha Yasseri[1,2]

[1] Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK,
[2] Alan Turing Institute, London NW1 2DB, UK,
taha.yasseri@oii.ox.ac.uk

Summary. Football is a major sport with worldwide popularity. In recent years excessive monetization of the game has been argued to have affected the quality of the match in different ways. In this work we take a data-heavy network science approach to measure predictability of football over 26 years in major European leagues. We benchmark our model against betting house predictions and after establishing its robustness we show that over time, the games in major leagues have become more predictable. We provide further support for this observation by showing that inequality between teams has increased in accord with the trends in predictability and the home field advantage has been vanishing ubiquitously. This is a first attempt to study football at a large scale and within a historical framework.

1 Introduction and Model

Football is worthy of extensive studies, as it is played by roughly 250 million players in over 200 countries and dependencies, making it the world's most popular sport [1]. A major question in relation to such a massive entertainment enterprise is if it can retain its attractiveness through surprise element or, due to significant recent monetizations, it is becoming more predictable and hence at the risk of losing popularity?

There has previously been a fair amount of research in statistical modelling and forecasting in relation to football [2–4]. A rather new approach in predicting performance is based on machine learning and network science. Most of the past research in this area however either focuses on inter-team interactions and modelling player behaviour rather than league tournament's results prediction, or are limited in scope—particularly they rarely take a historical approach in order to study the game as an evolving phenomenon. In the present work we use a network science approach to quantify predictability of football in a simple and robust way and by calculating the measures in 26 years of 11 major European leagues we examine if predictability of football has changed over time.

We build a directed network of all the matches within the training window, in which the edges point from the loser to the winner, weighted by the amount of points the winner earned. In the next step, we calculate the network *eigenvector centrality* score for all the teams. The recursive definition of eigenvector centrality, that is that the score of each node depends on the score of its neighbours that send a link to it, perfectly solves the problem of the dyadic scoring system mentioned above [5]. An example of such network and calculated scores are presented in Figure 1. We can calculate the score

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

difference between the two competing teams for any match after the Nth match. We will have $(T-N)$ matches with their respective outcomes and *score differences*. Finally, the logistic regression model will provide the probabilistic assessment of the score system for each match, allowing us to understand how correctly the outcomes are being split as a function of the pre-match score difference.

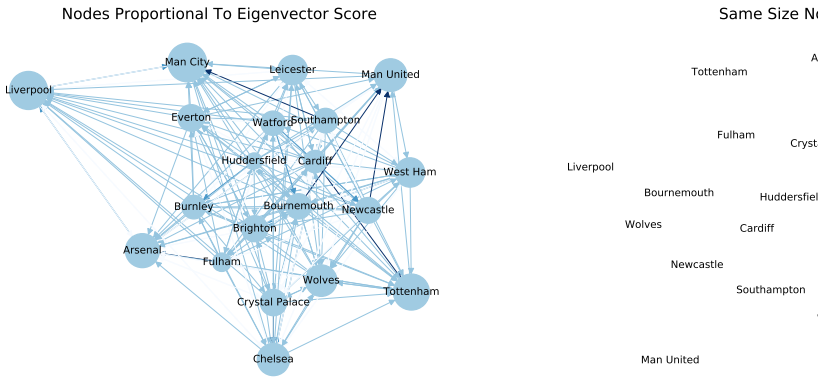


Fig. 1. The network diagram of the 2018-2019 English Premier League after 240 matches have been played (calculating centrality scores based on the last 190 matches).

2 Results and Conclusion

A positive trend in predictability is observed in most of the cases; See Figure. 2(left).

Increasing Inequality: In analysing the predictability of different leagues, we observe that predictability has been increasing for the richer leagues in Europe, whereas the set for which the indicator is deteriorating is composed mainly by peripheral leagues. It seems football as a sport is emulating society in its somewhat "gentrification" process, i.e., in richer leagues ricer teams are becoming even richer and stronger. We calculate the Gini coefficient of a given league-season's distribution of points each team had at the end of the tournament. Figure. 2(middle) depicts the values for all the leagues in the database, comparing the evolution of predictability and the evolution of inequality between teams for each case. The middles panels of Figure. 2 closely resemble the trends in left panels.

Home Advantage and Predictability: We calculate the home advantage from historical data by counting the total number of points that the home and away teams gained in each season. The trends in the share of home teams for different leagues are shown in Figure. 2(right). It is clear that the home field advantage is still present, however it has been decreasing throughout time for all the leagues under study. Increase in the number of foreign players, diminishing the effects of territoriality and its psychological factors, as well as observing that fewer people are going to stadiums, traveling is becoming easier, teams are camping in different pitches and players are accruing more international experience, can explain the reported trend; stronger (richer) teams are much more likely to win, it matters less where they play.

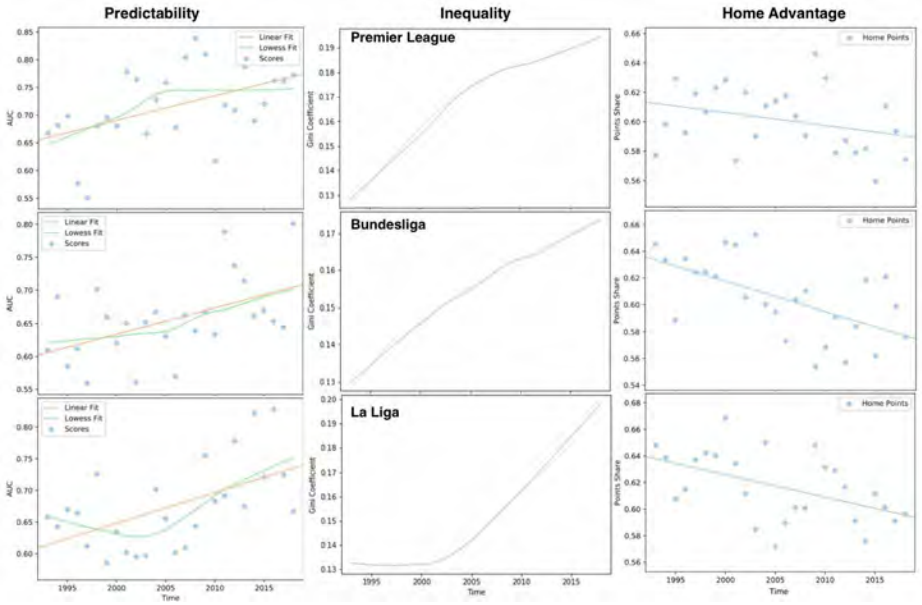


Fig. 2. Predictability (left), inequality (middle), and home advantage (right) for three major football leagues over time.

In conclusion: Relying on large-scale historical records of 11 major football leagues, we have shown that, throughout time, football is dramatically changing; the sport is becoming more predictable; teams are becoming increasingly unequal; and home field advantage is steadily and consistently decreasing for all the leagues in the sample. Future work should, as speculated in this work, try and assess: the role money is playing in removing the surprise element of the sport; expanding the sample barriers beyond the European continent; and ultimately, but not exhaustively, should test the money impact over predictability on different sports and leagues that – theoretically – should not be affected by it, namely leagues and sports that impose salary caps over their teams, such as the United States of America's Professional Basketball League (the NBA).

References

1. Giulianotti, R., Rollin, J., Weil, E., Joy, B., Alegi, P.: Football. encyclopedia britannica (2017)
2. Rue, H., Salvesen, O.: Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D (The Statistician) **49**(3) (2000) 399–418
3. Crowder, M., Dixon, M., Ledford, A., Robinson, M.: Dynamic modelling and prediction of english football league matches for betting. Journal of the Royal Statistical Society: Series D (The Statistician) **51**(2) (2002) 157–168
4. Constantinou, A.C., Fenton, N.E., Neil, M.: pi-football: A bayesian network model for forecasting association football match outcomes. Knowledge-Based Systems **36** (2012) 322–339
5. Newman, M.: Networks. Oxford university press (2018)

The Evolution of Digital Technologies: A Network Perspective on Machine Learning

Fabian Braesemann[1]*

[1]University of Oxford, Saïd Business School, Oxford OX1 1HP, UK,
fabian.braesemann@sbs.ox.ac.uk

1 Introduction

The rapid technological development due to innovations in Information and Communication Technologies brings vast economic opportunities. At the same time, it might lead to major shifts in the labour market due to technology-enabled offshoring or automatisation of jobs [1]. In particular, digital technologies such as 'Machine Learning' are predicted to have a profound impact on the economy [2]. However, their constituent parts and relations to other technologies are often ill-defined [3]. It is important for technology-specific investments and retraining programs to better understand their evolution and relation to other digital technologies.

Here, we construct a network of technologies related to machine learning based on data from *Stack Overflow*, the world's largest question-and-answer website for programming questions.[1] This network reveals the changing centrality of machine learning topics, libraries, and related programming languages over time as the network links rewire when novel technologies are introduced. It thus allows for understanding the development of the field as combinatorial technological evolution [4], shaped by the replacement of older technologies by novel ones. The data can be used to test network models on innovation and novelty [5, 6], and on creative destruction [7].

2 Data and Methods

Stack Overflow provides more than 18 million questions on thousands of different programming-related topics.[2] Most of these topics refer to technologies such as *Python*, *MATLAB* or to technology domains such as *Machine Learning*.[3] Each question is assigned one or more tags. Here, I focus on all questions tagged with the label 'machine-learning'. A question is represented as a binary vector containing a one, if tag A is present and zero otherwise. The total dataset contains $N = 119,926$ questions (rows) and $T = 2793$ tags (columns) posted between 2008 and 2019.

*ORCID-ID: 0000-0002-7671-1920

[1]It is presumed that it is feasible to represent the relations between digital technologies based on co-occurrences on the online platform.

[2]All Stack Overflow data are publicly available at https://archive.org/details/stackexchange.

[3]For simplicity, I assume that each tag refers to a technology; in its wider meaning as a 'a means to fulfill a purpose' [4]. Thus, I will use the terms 'tag' and 'technology' interchangeably.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

On this dataset, we applied Association Rule Learning [8] to construct a network at yearly intervals. The associaton rule concept *lift* is used, as it provides a balanced measure of proximity between two technologies (tags) in the *Technology Space*, similar to Hidalgo et al.'s [9] proximity measure of products in the product space. Formally, the lift between two technologies A and B is their joint ocurrence probability divided by the technologies' unconditional probabilities:

$$\text{lift}_{A,B} = \frac{P(E_A \cap E_B)}{P(E_A)P(E_B)},$$

where E_A and E_B are the events that questions refer to technology A and B, respectively. A lift > 1 implies that two technologies tend to occur together. Accordingly, this is the threshold for a link to be established between two technologies (nodes) in the network.

In the resulting yearly networks, I calculate the *normalised* betweenness centrality[4] of the individual technologies as a measure of their importance.

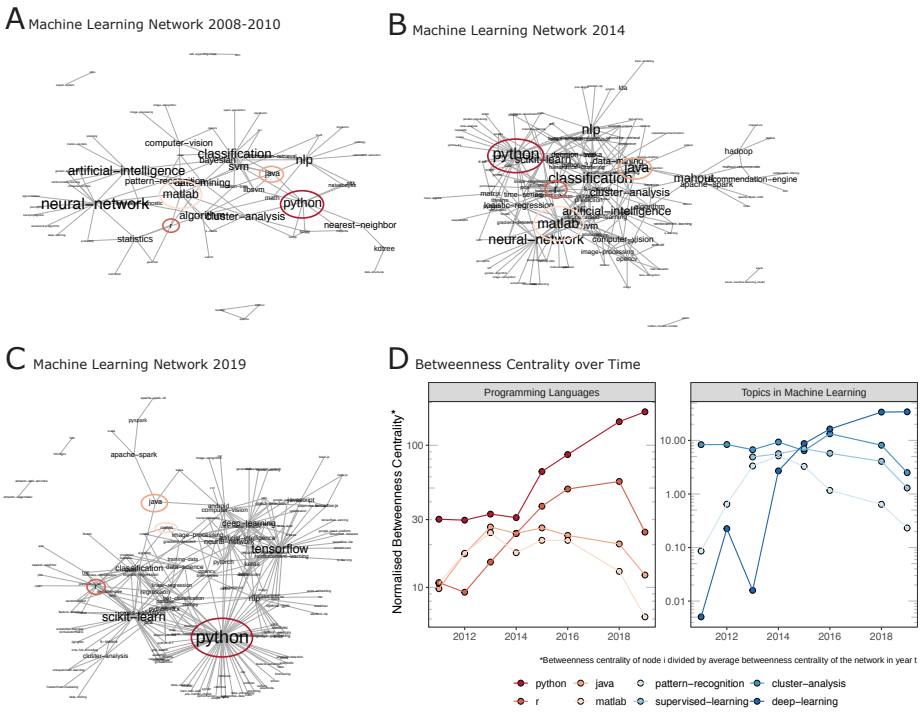


Fig. 1. (**A-C**) Networks of Stack Overflow tags related to 'Machine Learning' (ML) in 2008 – 2010, 2014, and 2019. Node size corresponds to betweenness centrality. The network became larger and denser over time as more ML-technologies are introduced. The centrality of four important programming languages, which can be used for ML, changes over time.
(**D**) Normalised betweenness centrality of four general programming languages (left panel) and four topics related to machine learning (right panel) from 2011 to 2019 (logarithmic scale). With the shifting focus from statistics to deep learning, Python's importance increased together with Python-based deep learning tools such as TensorFlow.

[4]The Betweenness centrality of the nodes is divided by the average betweenness centrality of all nodes in that year to allow comparison between networks.

3 Results

Figure 1 shows the network of technologies (tags) related to 'Machine Learning' in 2008 – 2010, 2014, and 2019 based on the Stack Overflow data. Four important programming languages (Python, R, Java, MATLAB), which can be used for machine learning applications, are highlighted by coloured circles. The early network in 2008 – 2010 is comparatively small and sparse, and the four programming languages have comparable positions in terms of their centrality. Within one decade, the set of technologies related to machine learning has changed considerably: Python has become the dominant programming language, closely related to the shift towards deep learning as a main paradigm in machine learning. Accordingly, Python benefited from the rise of Python-based deep-learning applications such as TensorFlow. The other languages have largely been displaced by Python within the domain of machine learning, due to its 'fitness' in generating productive 'offspring' technologies.

Summary. The development of digital technologies such as Machine Learning can be described empirically as a co-evolving network based on online platform data. Revealing the changing network relations is important to understand innovations in the digital sphere, as combinatorial possibilities between digital technologies are likely to be conditioned by their proximity in the *Technology Space*. The described network dataset provides a unique perspective on the technology space as it evolves in real-time. This perspective might help to better understand the geographical distribution of digital knowledge [10, 11]and innovations [12] in the digital sphere.

References

1. C. B. Frey, M. A. Osborne, The future of employment: How susceptible are jobs to computerisation?, *Technological forecasting and social change* 114, (2017) 254-280.
2. E. Brynjolfsson, T. Mitchell, What can machine learning do? Workforce implications, *Science* 358(6370), (2017) 1530-1534.
3. A. De Mauro, M. Greco, M. Grimaldi, P. Ritala, Human resources for Big Data professions: A systematic classification of job roles and required skill sets, *Information Processing Management* 54(5) (2018) 807-817.
4. B. Arthur, *The nature of technology: What it is and how it evolves*, Simon and Schuster, 2009.
5. F. Tria, V. Loreto, V. D. P. Servedio, S. H. Strogatz, The dynamics of correlated novelties, *Scientific reports* 4, (2014) 5890.
6. I. Iacopini, S. Milojevi, V. Latora, Network dynamics of innovation processes, *Physical Review Letters* 120, (2018) 048301
7. S. Thurner, P. Klimek, R. Hanel, Schumpeterian economic dynamics as a quantifiable model of evolution, *New Journal of Physics* 12(7), (2010) 075029.
8. R. Agrawal, T. Imieliski, A. Swami, Mining association rules between sets of items in large databases, in: *Acm sigmod record*, 22(2), (1993) 207-216, ACM.
9. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The Product space conditions the development of nations, *Science* 317(5837), (2007) 482-487
10. F. Stephany, F. Braesemann, An exploration of Wikipedia data as a measure of regional knowledge distribution, in: *International Conference on Social Informatics*, 10540 (2017), 31-40, Springer, Cham
11. F. Braesemann, N. Stoehr, M. Graham, Global networks in collaborative programming. *Regional Studies, Regional Science*, 6(1), (2019) 371-373.
12. F. Stephany, F. Braesemann, Coding together - coding alone: The role of trust in collaborative programming, in: *SocArxiv preprint*, 10.31235/osf.io/8rf2h, (2019)

Gender diversity in collaboration networks and the online popularity of scientists

Orsolya Vásárhelyi[1], Igor Zakhlebin[2], Staša Milojević[3] and Emőke-Ágnes Horvát[2]

[1] Department of Network and Data Science, Central European University, Budapest, Hungary
[2] Department of Communication Studies Northwestern University, Evasnton, IL, USA
[3] Center for Complex Networks and Systems Research, School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA

1 Introduction

Despite improvements scholarly activity is poised with gender-related inequalities. For example, female scientists still earn less [1], have access to less funding [2], are less likely to become full professors [3], their work receives fewer citations [4], and they benefit less from co-authorship [5]. These disparities persist, notwithstanding the evidence that female members increase the overall intelligence of teams [6] and that gender heterogeneous scientific teams are more creative and produce higher quality science [7].

It is believed that science dissemination is the crucial first step in exposing scholars' work to other scientists and the public. It is unclear, however, whether the online sharing of scientific articles mitigates, perpetuates, or reinforces inequalities that exist offline between male and female scholars. Since its inception in 2010, Altmetric.com has been heralded as an appropriate set of data sources and indicators to capture early interactions of scientific and lay audiences with scholarly output. Higher Altmetric scores have been shown to correlate with the number of citations [8], but little is known about potential inequalities in the online coverage of female scholars' work. Therefore, our main research question is: *Which dimensions of gender inequalities characterize science dissemination online?*

We studied differences in the dissemination of the articles of 371,800 scientists who had at least one article shared online in 2012. To identify factors associated with differences in dissemination, we collected meta-data about these scientists and their scholarship. In particular, we gathered their publication history and collaboration network for the five preceding years from the Open Academic Graph [9]. We also used Web of Science data [4] to determine scientific fields based on the references of individual publications [10] and to generate topics using article titles [11]. To investigate gender's effect on science dissemination, we inferred author's gender with a method based on their first names [12]. The used gender inference algorithm handles international names well and yielded 58% men, 25% women, and 17% unknowns among the considered scientists.

Here we present two main parts of our results: First, we show differences in the online coverage of research fields and topics, and their connection with gender. Then, using regression models, we highlight the importance of gender-related characteristics of collaboration networks in determining online popularity.

[4] This work uses Web of Science data by Clarivate Analytics provided by the Indiana University Network Science Institute and the Cyberinfrastructure for Network Science Center at Indiana University.

2 Results

Inequalities in online popularity across fields and topics We found striking inequalities in the dissemination of scientific articles based on research fields and topics. Out of the articles registered on Web of Science those belonging to fields such as Psychology and Medical Sciences received the most coverage (quantified by the number of shares), while Engineering and Social Sciences are disseminated least. Similarly, when we correlated the number of articles written overall about a certain topic with the number of shares of articles about the same topic, we found that in fields with a higher ratio of women such as Medical Sciences and Psychology, co-author teams where the fraction of female scholars was higher than the median percentage of females in the field, tend to be better at selecting topics with a larger online popularity. In Medical Sciences and Psychology, for example, articles written by female-majority co-author teams correlated most with number of article shares online (ρ=0.65 and ρ=0.45, respectively). Note that majority here is defined in comparison with the median female representation in the field, meaning that a three-author team containing one female scholar would be considered a female-majority team in e.g. Engineering and Physics. The most highly shared topics that female-majority teams write about are medical records, radiation, childhood leukemia and chronic pain in Medical Sciences, and aggression, autism, and the marshmallow test of delayed gratification in Psychology. In Mathematics and Computer Science only diverse teams topics correlated positively with the number of article shares online. Popular topics that diverse teams in Computer Science study are social media mining and sentiment analysis, while diverse Mathematician teams focus on high-ordered curved mesh generation and graphs. Despite these trends, women benefit less from being part of a team that publishes on popular topics as they are clearly under-represented among the scholars with highest online coverage (Figure 1). Accordingly, while there are differences between fields (e.g. we found fewer popular female Physicists online than Medical Scientists), there are less women among the most highly shared scholars across the board even when compared with their ratio in the population of scientists whose work is shared online. This trend holds both at the level of the top 5% scholars (superstars based on article shares) and top 25% (popular scientists).

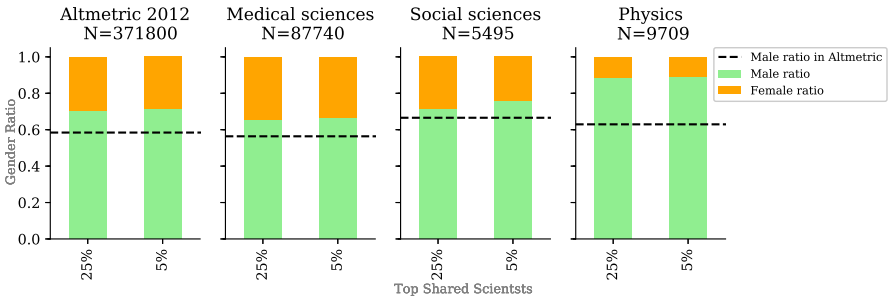


Fig. 1. Ratio of women among the top 25 and 5% of the most popular scientists by research field. Dashed lines indicate the ratio of men among all authors who published in the given field and were mentioned on Altmetric in 2012.

Predicting online popularity Both at the level of superstars and popular scientist we predict popularity as quantified by online coverage using logistic regression models. Interestingly, we can predict male scholars' popularity with higher precision than female scholars' popularity (top 25% : $AUC_{women} = 0.866$ and $AUC_{men} = 0.89$; top 5%: $AUC_{women} = 0.86$ and $AUC_{men} = 0.88$), which indicates a clearer behavioral template for men than women to achieve success in this context. Our models show that productivity (number of articles published in the preceding 5 years)

and research fields are the basis of online popularity. For superstar women (top 5%), productivity is more important than for men (Figure 2). The same factors matter for popular and superstar scientists, but the effect sizes in these groups are different for the two genders. Both men and women are negatively affected by the lack of gender-diversity in their ego networks (*male and female homophily in ego network*), but benefit from having same-gender collaborators *(ratio of male and female co-authors)*. Due to power-relations, similarly to others' results in a different knowledge production context [14], we found that men tend to become less popular when working with women. Smaller co-author teams *(average number of co-authors in the last 5 years)* are associated with higher online popularity for both genders, but dense ego networks (brotherhood or sisterhood-like collaborations) penalize only women.

We presented evidence that the gender portfolio of scientists' collaboration networks predicts online popularity, which makes it harder to overcome gender inequalities that exist offline among scholars. Men are over-represented among popular scientists, although diverse and female-majority research teams pick more popular topics. Our results also suggest less variance in the factors associated with male scientists' online popularity. The most important predictors of success are still based on academic merit and productivity, but gendered collaboration tie-formation differentiate men and women: successful scientists in social media, are embedded into opposite-gender academic circles, while maintaining same-gender personal relations.

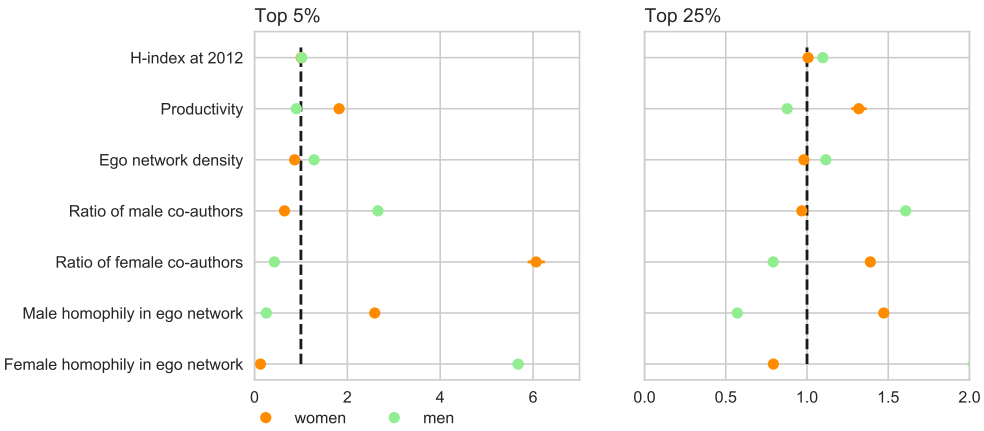


Fig. 2. Odds Ratios of selected variables for top 25% (popular) and top 5% (superstar) scientists based on article shares online. Models were ran separately for men and women and contained additional controls for fields, paper attributes, and publication history.

References

1. Shen, H.: Inequality quantified: Mind the gender gap. Nature. 495, 22 – 25 (2013)
2. Ley, T. J., Hamilton, B. H: The Gender Gap in NIH Grant Applications. iv. Science. 422, 1472–1474 (2008)
3. MossRacusin, C. A. , Dovidio, J. F., Brescoll V. L., Graham, M. J., Handelsman, J. .: Random plane networks. J. PNAS. 109(49), 16474–16479 (2012) (1961)
4. Lariviere, V., Ni, C., Gingras Y., Cronin, B., Sugimoto, C. R.: Bibliometrics: Global gender disparities in science. Nature. 504, 211–213. (2013)
5. Sarson, H.:Recognition for Group Work: Gender Differences in Academia. American Economic Review 107(5) (2017)
6. Bear, J.B., Williams Wolley A.: The role of gender in team collaboration and performance. Interdisciplinary Science Reviews. 36, 146–153 (2011) 423 (2006)
7. Campbell, G., Mehtani, S., Dozier, M.E., Rinehart, J.: Gender-Heterogeneous Working Groups Produce Higher Quality Science. PLoS ONE. 8(10), (2013)
8. Thelwall, M., Haustein, S., Larivier, V. Sugimoto, C. R.: Do Altmetrics Work? Twitter and Ten Other Social Web Service. PLoS ONE. 8(5), (2013)
9. Tang, J., Zhang,J., Yao,L., Li,J., Zhang,L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD2008). pp.990-998.
10. Milojević, S.: Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. Quantitative Science Studies. (under revision)
11. Milojević, S.: Quantifying the cognitive extent of science. Journal of Informetrics. 9(4), 962–973 (2015)
12. Ford, D., Harkins, A., Parnin, C.: Someone like me: How does peer parity influence participation of women on stack overflow?. 2017 IEEE Symposium on Visual Languages and Human-Centric Computing. (2017)

13. Lutter, M.: Do Women Suffer from Network Closure? The Moderating Effect of Social Capital on Gender Inequality in a Project-Based Labor Market, 1929 to 2010. American Sociological Review. 80(2) 329–358 (2015)
14. Vedres, B., Vasarhelyi, O.: Gendered behavior as a disadvantage in open source software development. EPJ Data Science. 8(25), (2019)
15. Cullen, D., Luna, G.: Women Mentoring in Academe: addressing the gender gap in higher education. Gender and Education. 5(2) 125–137. (1993)
16. Bornmann, L., Daniel, H. D.: Does the h index for ranking of scientists really work?, Scientometrics. 65(391) (2005)
17. Barnes, C.: The h-index Debate: An Introduction for Librarians. The Journal of Academic Librarianship. 43(6), 487–494 (2017)
18. Fox, C. W., Timothy Paine C. E.: Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. Ecology and Evolution. 9, 3599–3619 (2019)

The discriminative power of online social networks [1]

Xiaoyan Qiu[1,2], Diego F. M. Oliveira[2], Alireza Sahami Shirazi[3], Alessandro Flammini[1], and Filippo Menczer[2,3]

[1] School of Economics and Management, Shanghai Institute of Technology
[2] School of Informatics and Computing, Indiana University
[3] Yahoo Research
[4] Indiana University Network Science Institute

1 Introduction

Social media are massive marketplaces where ideas and news compete for our attention [2]. Previous studies have shown that quality is not a necessary condition for online virality [3] and that knowledge about peer choices can distort the relationship between quality and popularity [4]. We investigate [1] quality discrimination in a stylized model of online social network, where individual agents prefer quality information, but have behavioral limitations in managing a heavy flow of information. We measure the relationship between the quality of an idea and its likelihood to become prevalent at the system level. We find that both information overload and limited attention contribute to a degradation in the market's discriminative power. A good tradeoff between discriminative power and diversity of information is possible according to the model.

2 A model for choices in a networked environment

We aim to examine the conditions in which the "best" ideas are those that capture a greater portion of collective attention, and whether this happens at the expense of the diversity of ideas. To this end, we propose a simple agent-based model inspired by the long tradition of representing the spread of ideas as an epidemic process where messages are passed along the edges of a network [5]. Agents are represented by the nodes of a static network where the links embody social connections. The network dynamics in the model capture the salient ingredients common to popular social media platforms. Each message, or post, carries a "meme" or "idea," i.e., the unit of information that spreads from person to person [6]. Different messages may carry the same meme.

We imagine that each meme is characterized by an intrinsic *quality* value. We assume that the probability that an agent shares one of these memes, allowing it to spread, is proportional to the meme's quality. The quality might represent different properties that make the meme more likely to be shared: the originality of an idea, the beauty of a picture, and the truthfulness of a claim are valid examples. Messages carrying new memes are continuously introduced into the system in an exogenous fashion at a rate rate μ, a parameter that measures the *information load* of the agents.

Agents produce messages containing new memes and reshare messages originated or forwarded by their neighbors. When resharing, an agent is capable of paying *attention* to only a finite number α of messages at a time. If we think of messages from

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

neighbors as appearing in, say, reverse chronological order on a social media feed, a user during a session will scroll down the feed to view α recent posts. Further details about the model are presented in [1].

3 Results

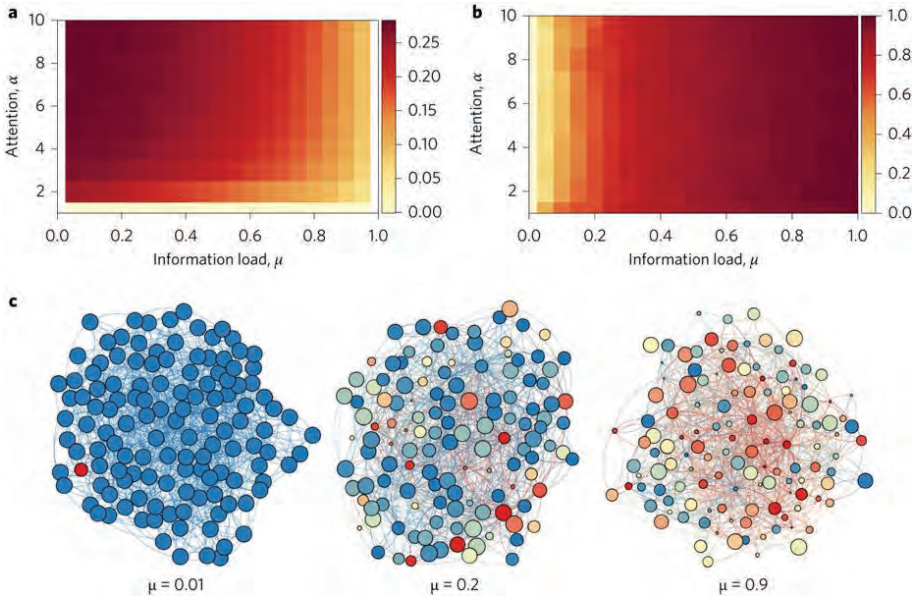


Fig. 1. a, Discriminative power τ (colour scale bar) as a function of information load and finite attention. b, Diversity H/H($\mu = 1$) (colour scale bar) as a function of intensity of information load and attention. c, Illustration of the tradeoff between the discriminative power of the system in spreading quality memes and diversity of content in the network. Nodes represent agents, their colour represents the last shared meme and their size indicates the quality of that meme (the bigger the node, the higher the quality). Edges represent social connections among agents, such as followers on Twitter or friends on Facebook. When the information load μ is small, only high-quality memes are present, with low diversity. As μ increases we observe higher diversity and lower discriminative power. Here N=128 and α=10.

We can summarize the dependency between the quality of memes and their success in a single *discriminative power* measure by looking at the correlation between quality and popularity, defined as the number of reshares a given meme has received during its lifetime. We employ the Kendall rank correlation coefficient τ [7]. High τ indicates that fitter memes are more likely to win, granting the system discriminative power; in the extreme case $\tau = 1$ the two rankings are completely concordant. Small τ signifies a lack of quality discrimination by the network.

Discriminative power in spreading quality content is a desirable property of a social network. A second desirable property of an ideal communication system is the preser-

vation of information diversity, i.e., the possibility to have many distinct memes alive simultaneously. To measure the amount of diversity in the system at the steady state, we start from the entropy $H = -\sum_m P(m) \log P(m)$ where $P(m)$ is the portion of attention received by meme m, i.e., the fraction of messages with m across all of the user feeds. The sum runs over all memes present at a given time and is averaged over a long period after stationarity has been achieved. The minimum entropy is zero, when all nodes have the same meme ($\mu = 0$). The maximum entropy, obtained in the extreme case $\mu = 1$, depends on α. Discriminative power and diversity are in contradiction — the price associated with the capability of the network to let a high-quality meme prevail is a loss in diversity, with many memes receiving relatively small attention despite their intrinsic quality.

The tradeoffs between discriminative power and diversity is illustrated in Fig. 1. For any value of finite attention $\alpha > 1$ we observe a transition from relatively high discriminative power and low diversity (when information load is low) to high diversity and low discriminative power (high information load). The amount of attention α has a significantly effect on the tradeoff: for a given level of diversity the discriminative power improves when people can pay attention to multiple memes, and vice versa the network can sustain a larger diversity without loss in discriminative power. When α is large, there is a region where the network can sustain very high diversity with relatively small loss in discriminative power.

The proposed model is quite minimal and relies on few parameters, but it captures salient behavioral features that shape the diffusion of information in online social networks. This allows us to study how information load and limited attention affect the discriminative power of the network, i.e., the likelihood that the best memes will succeed at reaching many people. Our main finding is that the survival of the fittest is far from a foregone conclusion where information is concerned. Both information load and limited attention lead to low discriminative power, so that it becomes very difficult for the best memes to win. Meme diversity can coexist with network discriminative power when we have plenty of attention and are not overloaded with information. For a full account, please see [1].

References

1. The present paper is derived by: Qiu, X., Olivera, D. F. M., Shirazi A. F., Flammini, A., Menczer F.: Limited individual attention and online virality of low-quality information. Nat. Hum. Beh. 1, 0132 (2017).
2. Simon, H.:Designing Organizations for an Information-Rich World, in Computers, Communication, and the Public Interest, 37–52, The Johns Hopkins Press (1971).
3. Weng, L., Flammini, A., Vespignani, A. and Menczer, F.: Competition among memes in a world with limited attention. Sci. Rep. 2, 335 (2012).
4. Salganik, M.J. and Dodds, P. S. and Watts, D. J.: Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science, 854–856, 311 (2006).
5. Bailey, N. T. J.: The mathematical theory of infectious diseases and its applications. Charles Griffin & Co. (1975).
6. Dawkins, R.: The selfish gene. Oxford University Press (1989).
7. Kendall, M.: A New Measure of Rank Correlation. Biometrika 30, 81–89 (1938).

Relational Identities: a graph-theoretical approach with applications to online social media.

Emanuele Cozzo[1,2,3], Luce Prignano[1,2] Antonio Calleja-Lopez[3], and Albert Diaz-Guilera[1,2]

[1] Departament de Fsica de la Matria Condensada, Universitat de Barcelona
[2] Universitat de Barcelona Institute of Complex Systems-UBICS
[3] Communication Networks and Social Change-Internet Interdisciplinary Institute of the Open University of Catalonia

1 Introduction

The representation and the construction of the self has been thought to involve multiple components since the early days of *self* research in psychology and social-cognitive science [1]. Researches suggested that there exist three self-aspects [2]: individual, relational, and social. Moreover, different self-aspects form distinct cognitive structures. The individual identity refers to the conception of oneself as autonomous and unique, having a clear boundary from others. Social identity refers to the self-definitions derived from one's membership in groups or social categories. Here we focus on the the relational identity, which refers to aspects of the self associated with one's relationships with *significant others*. In the tradition of social network analysis, this aspect is associated with the study of ego-networks [3]. Dunbar and collaborators [4, 5] showed that alters in ego-networks are organized in different layers of decreasing strength in relation with the ego, with a strong inner circle of more or less 5 persons, and that the structure of online social networks mirrors those in the offline world [6]. From the point of view of social psychology, in the late 90s, Aron et al. [7] proposed the *Including others in the self* model for the relational self. This model proposes that, to some extent, people treat the resources, perspectives, and identities of significant others as their own. By combining these two perspectives, in this work we go beyond ego-networks and propose the definition of *relational identity graph* (RIG) as a graph-theoretical operationalization of the concept of collective identity emerging from relational self-construals [1]. To do so, we need to build new instruments in the realm of graph theory, expanding some classical definitions and algorithms. Finally, we show how the internal structure of the relational identity can be characterized by means of graph metrics, with special attention to metrics that quantify social identity complexity [8].

2 Relational Identity Graphs

Definition 1. *Given a set of social actors V, define:*

(a) a social function $H(x,y) : V \times V \to \mathbb{R}$
(b) a relevant condition $\mathbf{r}$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

(c) a binary relevant relation: $\mathbf{R} : (x,y) \in \mathbf{R}$ *if and only if* $H(x,y)$ *satisfies* $\mathbf{r}$.

The social function has to provide a measure of the strength of the relation between two social actors in V, while the relevant condition defines a threshold, in general dependent on x, above which the relation is considered relevant for x. Thus, $\mathbf{R}$ defines a directed graph such that there is an edge from social actor x to social actor y if y is relevant for x, with respect to the function H and the relevant condition $\mathbf{r}$. In general the social function H has not to be symmetric, i.e. $H(x,y) \neq H(y,x)$ and thus the graph associated to $\mathbf{R}$ is directed. Note also that, even in the case $H(x,y) = H(y,x)$ this does not necessarily imply that both satisfy the relevant condition, and therefore the result may still be a directed graph. From a graph-theoretical point of view, a social function H defines by itself a weighted graph and the graph associated to $\mathbf{R}$ is an edge-induced subgraph of it. Thus, $\mathbf{R}$ can be interpreted as the backbone of a social network with weights $H(x,y)$ [9].

Let's define the relevant neighbourhood of a social actor as the operationalization of the concept of significant others.

Definition 2. *Consider a social actor x, we call the relevant neighbourhood of x in V the set* $\mathbf{R}(x) = \{y : (x,y) \in \mathbf{R}\}$

That is, the relevant neighbourhood $\mathbf{R}(x)$ is the out-neighbourhood of x in the graph G associated to $\mathbf{R}$. We are now able to give the definition of the Relational Identity Graph of a social actor.

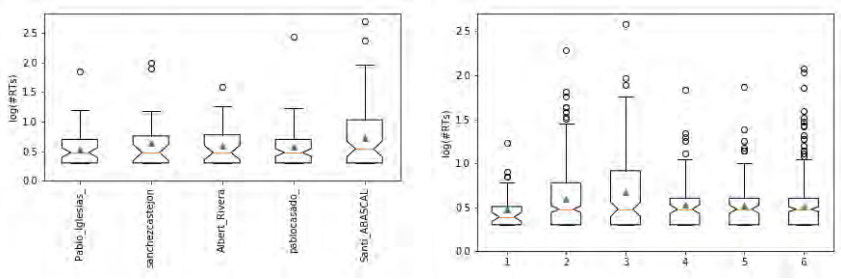
Definition 3. *We call the Relational Identity Graph of x w.r.t.* $\mathbf{R}$*,* $RIG(x)$*, the maximal strongly connected component of the vertex-induced subgraph* $G(S(x))$ *that includes x, where* $S = \mathbf{R}(x) \cup \mathbf{R}(y) \cup \mathbf{R}(z), \forall y \in \mathbf{r}(x)$ *and* $\forall z \in \mathbf{R}(y)$.

A $RIG(x)$ operationalizes the concept of the relational identity of a social actor. Building on graph-theoretical concept, a number of structural properties of a $RIG(x)$ can be defined and related to the relational identity. We can give a generalizations of the concept of separator [10] from undirected to directed graphs. Minimal strong separators are of great importance in the contest of RIGs since they are the minimal set of social actors in the relational identity that can break the identity and form a new one, and they are directly related to social identity complexity.

3 Results: On-line Political Relational Identities

We study the RIGs of the leaders of major Spanish political parties and of political activists on Twitter. The multiple components of the representation of the self can be recognized also when looking to online identities. The bio, the profile picture, and other elements always present in any online social platform, are a representation of the individual self. Relational identity is represented in the interaction patterns. Focusing on Twitter, three types of interactions are possible: mentions (a user mentions another user in her post), likes (an expression of appreciations to another user post), and retweets (the user reposts another user post). The including others in the self model, i.e. the hypotesis that people treat the resources, perspectives, and identities of significant others as their

own, is just one of the proposed mechanism of relational identification and there is still a lot of debate around it. However, when looking to online social self, the proposed mechanism is much more natural. Actually, sharing others posts complitely describes the including others in the self mechanism, like retweetting on Twitter. For this reason, we will focus on retweets.



(a) (b)

Fig. 1. Retweets statistics of major party leaders -panel (a)- and political activists -panel (b)-.

We define the relevant condition as being an outlier in the retweet statistics,i.e. accounts that recive a number of retweets that exceeds the third quartile by more than 1.5 times the interquartile range. The distribution of such outliers is quite different (see Fig.1) for party leaders and political activists. We construct the RIG of each actor and compare their characteristics.

References

1. Brewer, M. B., & Gardner, W. (1996). Who is this" We"? Levels of collective identity and self representations. Journal of personality and social psychology, 71(1), 83.
2. Sedikides, C., & Brewer, M. B. (2015). Individual self, relational self, collective self. Psychology Press.
3. Freeman, L. C. (1982). Centered graphs and the structure of ego networks. Mathematical Social Sciences, 3(3), 291-304.
4. Dunbar, R. I. (2014). The social brain: Psychological underpinnings and implications for the structure of organizations. Current Directions in Psychological Science, 23(2), 109-114.
5. Tamarit, I., Cuesta, J. A., Dunbar, R. I., & Snchez, A. (2018). Cognitive resource allocation determines the organization of personal networks. Proceedings of the National Academy of Sciences, 115(33), 8316-8321.
6. Dunbar, R. I., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. Social networks, 43, 39-47.
7. Aron, A., McLaughlin-Volpe, T., Mashek, D., Lewandowski, G., Wright, S. C., & Aron, E. N. (2004). Including others in the self. European review of social psychology, 15(1), 101-132.
8. Roccas, S., & Brewer, M. B. (2002). Social identity complexity. Personality and Social Psychology Review, 6(2), 88-106.
9. Serrano, M. ., Bogun, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. Proceedings of the national academy of sciences, 106(16), 6483-6488.

10. Tarjan, R. E. (1985). Decomposition by clique separators. Discrete mathematics, 55(2), 221-232.

Part XVII

Resilience and Control

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Transporters: Spring-systems in disguise. A physics model for analysing transporter networks.

Jonathan Bourne[1]

University College London, Gower street, UK,
jonathan.bourne.15@ucl.ac.uk

1 Introduction

Transport networks are systems that move resources from the point where they are generated to the point where they are demanded or consumed, and are ubiquitous in modern society. this class of networks include the physical transport network moving people and goods but also other network types such as telecommunications systems moving information, or utilities networks such as gas or electricity. One of the greatest issues facing transport networks is that of cascading failures [8] caused by random events or targeted attack. A popular metric used in research relating to the robustness of a network is system tolerance (α) [8, 7]. The tolerance of line i is given by $f_i^{\max} = \alpha_i |f_i^c|$ where f_i^c is the power flow over line i under initial conditions, and α_i is the tolerance of line i. The system tolerance $\alpha = \frac{1}{n}\sum_{i=1}^{n} a_i$, where n is the total number of lines, can be used as a proxy for network robustness. However, system tolerance does not take account of the network's topological structure or the distribution of line tolerances. The work presented here combines line tolerance and the topological structure of the network into a single intuitive metric.

The core idea of this study is to re-represent a transport network as a system of springs, in order to provide insight into the network's robustness under attack. Spring systems are a useful physics model for explaining the complex relationships found in networks [4]. In this paper, the spring system begins as a 2D network, where the nodes are constrained to only move perpendicularly to the plane. Generation exerts an upwards force while demand exerts a downward force. As the generation and demand are equal, the system finds a 3D equilibrium when the forces generated by the nodes are balanced by the restorative force of the springs. The generation and demand at the nodes is converted to force using an arbitrary constant. The stiffness of the springs is calculated using $k = r(1 - \frac{1}{\alpha}) + c$ where c and r are arbitrary constants that define the range of stiffness k.

In this paper two key metrics are derived from the spring system; the height embedding of the nodes relative to each other represents the amount of generation support in that part of the network; the mechanical strain of the springs represents the robustness of the network. Strain, a common measurement within engineering, is defined as $\varepsilon = \frac{\Delta H - D}{D}$, where ΔH is the extension of the spring under a tensile force, with D representing its original length. The height and strain embeddings are found using an algorithm based on the equations of motion. The relationship between the nodes is subject to the n-body problem [3], and so the equilibrium is found through iteration, a

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

feature of this class of network algorithm [6]. The force experienced by each node is used to find the acceleration, velocity and distance for each time step Δt. The equation $z = v_{t-1}t + \frac{1}{2}a_t t^2 + z_{t-1}$ calculates displacement from the origin, where the velocity is given by $v_t = v_{t-1} + \frac{F_{\text{net},t}}{m}t$ and the acceleration by $a = \frac{F_{\text{net}}}{m}$. t, m, and F_{net} represent time, arbitrary mass and net force acting on the node, respectively. A damping factor is also included to ensure that the system loses energy and converges.

This paper demonstrates the value of the embeddings created by the spring system using IEEE-118 [2], the UK high voltage power grid generated from the ETYS dataset [1], and five non-flow networks demonstrated in [9].

2 Results

Using a 4 node toy network, I demonstrate that strain can express differences between networks in a way that system tolerance α cannot. In the toy example strain varies by up to 72% whilst α and the total system flow capacity remain constant.

I attack the IEEE-118 network until collapse under different load/generator profiles, using a DC power flow simulator [5]. For each load profile I set α to one of twelve values, then permute the excess capacity of the lines to generate 480 unique system α's, I then perform 100 simulations where I randomly attack the network to collapse, on each unique α. This creates 48,000 attack to failure simulations per load profile. I predict the point at which the network collapses using a loess model [10] where the independent variable is either strain or α. I find that strain ($R^2 = 0.97$) explains a greater proportion of the variance of the collapse point of the network than α ($R^2 = 0.92$), and RMSE is 38% lower for strain. This finding is consistent across load profiles.

I then find the strain of the UK power grid under base loading conditions and identify points of potential weakness, as shown in figure 1. In addition, visualising the UK high voltage network shows a North-South slope consistent with the movement of electricity on the UK mainland.

Finally I show that not only can the system be applied to non-flow networks, but strain does not suffer from the same issues as assortativity as described by [9] and can successfully distinguish the Peel's quintet. I do this by generating 100 examples from each graph class of the quintet, projecting the nodes into a space defined by mean node height and network strain. I train 100 logistic regression models, using 10 sets of 10-fold cross validation, the mean classification accuracy of the models is 97.6%. The accuracy is the result of the algorithm integrating topology and load and is despite the graph classes in the quintet having identical, assortativity, number of nodes, number of edges, and number of links between node classes.

Summary. This study demonstrates a novel approach to analysing the robustness of transport networks. It does this using a physics model that integrates the line capacities and the network topology into a single metric. The method provides a local and system-wide overview of embedded strain allowing the network to be tuned for robustness. The information produced can be intuitively interpreted by visualising it on a map. The method can also be applied to networks without flow or line capacity, even distinguishing between networks that have that have been designed to be identical using traditional network science metrics.

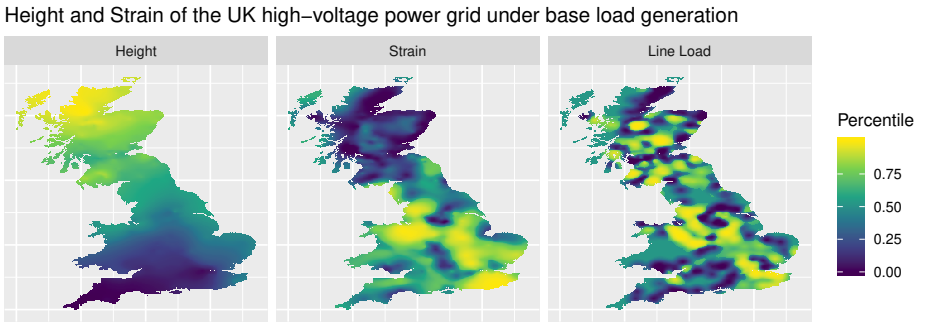


Fig. 1. The topology of Great Britain where height is the embedded height of the nodes as well as the strain and tolerance of the lines. Values at geographical points between nodes have been interpolated using kriging with spherical distance model

References

1. Electricity Ten Year Statement (ETYS) National Grid ESO
2. IEEE 118-Bus System - Illinois Center for a Smarter Electric Grid (ICSEG)
3. Aarseth, S.J.: The N-body problem. In: Gravitational N-Body Simulations: Tools and Algorithms, pp. 1–17. Cambridge Monographs on Mathematical Physics, Cambridge University Press (2003)
4. Bacco, C.D., Larremore, D.B., Moore, C.: A physical model for efficient ranking in networks. Science Advances 4(7), eaar8260 (jul 2018)
5. Bourne, J., O'Sullivan, A., Arcaute, E.: Don't go chasing artificial waterfalls: Simulating cascading failures in the power grid and the impact of artificial line-limit methods on results, https://arxiv.org/abs/1907.12848
6. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and Experience 21(11), 1129–1164 (1991)
7. Kinney, R., Crucitti, P., Albert, R., Latora, V.: Modeling cascading failures in the North American power grid. Eur. Phys. J. B 46, 101–107 (2005)
8. Motter, A.E., Lai, Y.C.: Cascade-based attacks on complex networks. Physical Review E 66(6) (2002)
9. Peel, L., Delvenne, J.C., Lambiotte, R.: Multiscale mixing patterns in networks. Proceedings of the National Academy of Sciences 115(16), 4057–4062 (apr 2018)
10. Savitzky, A., Golay, M.J.E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36(8), 1627–1639 (jul 1964), https://doi.org/10.1021/ac60214a047

Dimensions of stability in complex ecological networks

Virginia Domínguez-García[1], Vasilis Dakos[1], and Sonia Kéfi[1]

ISEM, CNRS, Univ. Montpellier, IRD, EPHE, Montpellier, France
virginia@onsager.ugr.es

1 Introduction

Stability has been a central topic of research in complex systems across disciplines. From financial systems to socioeconomic models of political regimes or genetic regulatory circuits, the study of dynamical stability keeps attracting the attention of the scientific community. This interest has been particularly prominent in ecology, where it has promoted decades of research [1–6]. Understanding the stability of ecological communities has become a matter of increasing importance in the context of global environmental change, yet, it has proved to be a challenging task. Multitude of metrics are used to assess the stability of ecological systems but different choices may result in conflicting conclusions. Estimating how many and which metrics need to be measured holds the key to improving our ability to evaluate the stability of natural systems. This can be achieved by studying how many 'dimensions' of stability there are, i.e. in how many independent components stability metrics can be grouped. We tackle this challenge from a theoretical perspective by investigating how stability metrics are intertwined in trophic ecological networks. Combining structural food-web models [7] with bioenergetic consumer-resource models [8, 9], we simulate the dynamics of multispecies trophic communities under three different perturbation scenarios: pulse perturbations (instantaneous changes in species' biomass after which the recovery o the community is studied), press perturbations (lasting changes after which the pre and post-perturbed communities are compared) and environmental stochasticity (continuous small external changes). We quantify the stability of our simulated communities to these perturbations with 27 metrics frequently used in the ecological literature, and build a network representation of the correlations of stability metrics in complex trophic communities, the 'stability network'. In this network representation, the nodes are the different metrics, and the links their unsigned spearman's correlation rank: the higher the value of the correlation between a pair of metrics, the thicker the link between them. By studying these correlations we can evaluate whether the different metrics considered provide similar information about the stability of an ecological community or whether they form distinct groups that reflect partly independent 'dimensions' of community stability.

2 Results

Applying an algorithm of community detection, based on maximizing the modularity, to the stability network revealed that stability metrics can be lumped into three main

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

groups of relatively independent stability components: 'early response to pulse', 'sensitivities to press' and 'distance to threshold'. Interestingly, the three emergent groups split metrics in terms of both the temporal scale of the response and the type of perturbation. Indeed, the 'early response to pulse' group only contains metrics describing transient behavior, while the 'sensitivities to press' and 'distance to threshold' groups contain metrics describing long-term (asymptotic) dynamics. Furthermore, the 'early response to pulse' and 'sensitivities to press' form two contrasting groups containing metrics that respectively refer to pulse and press perturbations, while metrics in the 'distance to threshold' group refer to both types of perturbations. Therefore, these three groups can be interpreted as different stability components that reflect different 'dimensions' of the stability of trophic communities [10], i.e. features that should be measured in an ecological community to accurately assess its stability. Selecting metrics from each of these groups allows a more accurate and comprehensive quantification of the stability of ecological communities.

Our results contribute to improving our understanding and assessment of stability in ecological communities. However, although our study focuses on the stability of food webs, the relationships found here could be of interest to understand the stability of other types of networks, in ecology as well as in other disciplines. After all, directed networks of many kinds describe transport of matter, information, or capital in a similar way as food webs describe fluxes of biomass from primary producers to apex predators. The framework we propose is flexible enough to accommodate to different conditions and opens a way towards simplifying the study of stability in any type of complex dynamical system.

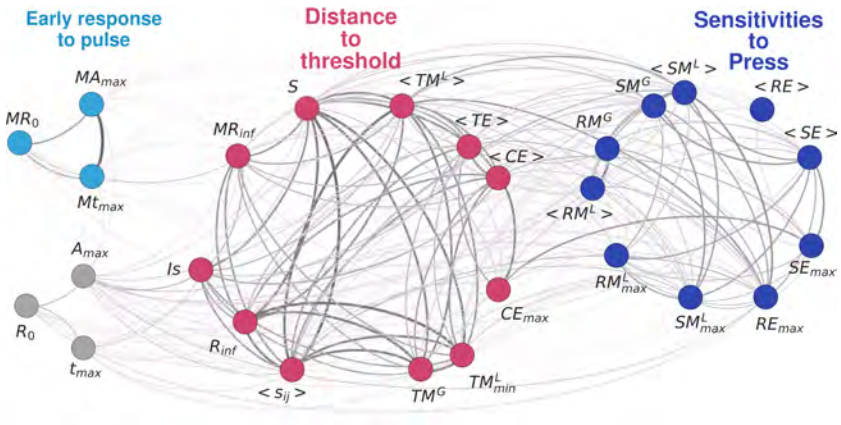


Fig. 1. Network of stability metrics (i.e. stability network) showing the three emergent groups of stability metrics. The nodes represent the different metrics used to quantify the stability of the trophic communities, and the links their unsigned spearman's rank correlation (the thicker the link, the stronger the correlation).

While the need to consider the multidimensionality of stability has been clearly stated in the ecological literature for decades, little is known about how different metrics of stability relate to each other in ecological communities. By simulating multispecies trophic networks, we measure how frequently-used stability metrics relate to each other. Using algorithms of community detection, we identify the independent components they form based on their correlations. Our results open a way to a simplification and better understanding of the overall stability of ecological systems.

References

1. Orians GH (1975) *Diversity, stability and maturity in natural ecosystems*, eds. van Dobben WH, Lowe-McConnell RH. (Springer Netherlands, Dordrecht), pp. 139–150.
2. Pimm SL (1984) The complexity and stability of ecosystems. *Nature* 307:321–326.
3. Grimm V, Wissel C (1997) Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. *Oecologia* 109(3):323–334.
4. Ives AR, Carpenter SR (2007) Stability and diversity of ecosystems. *Science* 317:58–62.
5. Donohue I, et al. (2016) Navigating the complexity of ecological stability. *Ecology Letters* 19(9):1172–1185.
6. Kéfi S, et al. (2019) Advancing our understanding of ecological stability. *Ecology Letters* 22(9):1349–1356.
7. Williams RJ, Martinez ND (2000) Simple rules yield complex food webs. *Nature* 404(6774):180–183.
8. Yodzis P, Innes S (1992) Body Size and Consumer-Resource Dynamics. *The American Naturalist* 139(6):1151–1175.
9. Brose U, Williams RJ, Martinez ND (2006) Allometric scaling enhances stability in complex food webs. *Ecology Letters* 9(11):1228–1236.
10. Donohue I, et al. (2013) On the dimensionality of ecological stability. *Ecology Letters* 16(4):421–429.

Optimizing Hospital Networks for Resource Allocation During a Large-Scale Disaster: A Sociotechnical Resilience Approach

Fredy Tantri[1,4], Cheung Sai Hung[2,4], and Sulfikar Amir[3,4]

[1] Interdisciplinary Graduate School, Nanyang Technological University, Singapore
fred0009@e.ntu.edu.sg
[2] School of Civil and Environmental Engineering, Nanyang Technological University
[3] Division of Sociology, School of Social Sciences, Nanyang Technological University
[4] Future Resilient Systems, Singapore-ETH Centre, Singapore

1 Introduction

The capability of healthcare as a sociotechnical systems to response to a crisis does not only depend on the logistics and physical infrastructure, but also on the informational relations to allow efficient coordination [1]. According to the Medical Surge Capacity and Capability (MSCC) Management System [2], there are six-tier of coordination to response a medical surge. Our research focus on the Tier 2, which is the coordination between hospitals. The benefit of having Tier 2 coordination is the rapid transfer of resource and information.

The aims of our research are to obtain the optimal networks of coordination between hospitals as a preparedness step to response to a large-scale disaster. At the same time, the model can be used as the decision support systems of the incident command center during the disaster by providing recommendations for resource allocation. We developed our network optimization method using stochastic approach and genetic algorithms.

2 Model and Algorithm Description

We used the hospital data in Jakarta which including the number of medical staff and hospital location to model our network and resources. We picked 14 random locations all over Jakarta as our disaster site to train our model. Node and edge respectively represent a hospital and Tier 2 coordination between hospitals. Each node has resources based on the number of available doctors. We assumed that the number of available doctors is 2/3 of the total doctors as the rest of them maintaining the public healthcare service of the hospital.

The genetic algorithm is used to obtain the optimal hospital networks. The algorithm involves the iteration of crossover, mutation, fitness evaluation and reproduction processes. During the fitness evaluation step, we calculate the performance of each Tier 2 Networks using Equation (1) and (2), which is calculated by taking the mean of the number of treated victims of all hospitals over a period of time T for M scenarios,

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where 4 out of 14 disaster sites are chosen randomly in each scenario Δ. Each disaster location causes V number of victims. They will be sent to all of the hospitals which are located within 15 minutes travel time. The victims are assumed to reach hospital h around disaster site in t_{dh} minutes and each one of them will be taken care by 1 doctor in 20 minutes. If a hospital i outside of disaster vicinity has a Tier 2 connection with a hospital j in disaster vicinity, it will send its resources to the affected hospital in $w_{edge(i,j)}$ minutes. The duration of t_{dh} and $w_{edge(i,j)}$ are based on the travel time data obtained using Google API. The maximum number of Tier 2 connection each hospital h can have is limited to two links. An edge can only be assigned between hospital i and hospital j if the travel time between them ($w_{edge(i,j)}$ or $w_{edge(j,i)}$) are less than W hours.

$$f(G) = \frac{\sum_{i=1}^{M} P(G|\Delta_i)}{M} \tag{1}$$

$$P(G|\Delta) = \sum_{h \in H} \sum_{t=1}^{T} p_h^t \tag{2}$$

and subjects to these following constraints,

$$|edges(h)| \leq 2 \tag{3}$$

$$min(w_{edge(i,j)}, w_{edge(j,i)}) \leq W \tag{4}$$

where, G is the evaluated hospital network (chromosome), H is the set of hospitals in the network, t is the simulation time in minute, and p_h^t is the number of treated victims in hospital h at time t

3 Results

We set the simulation number M to 500 scenarios. The T is set to 4 hours for each simulation because this model aims to improve the resilience during the very first few hours after a large-scale disaster where the condition is in chaos and the incident command center still needs to assess and coordinate all of the emergency-related agencies. The performance of the 1st, 34th, and 100th generation hospital networks are shown in Fig. 2. The shifting of the mean value to the right indicates the improvement of network performance. The generated hospital networks were tested by running a random disruption scenario to show how it allows the efficient resource sharing, indicated by the contour plot where the number of doctors are high (yellow to blue contour) in the affected hospitals (purple node) around disaster site (marker icon), while most of the other unaffected hospitals (green nodes) which send their resources are in red contour area. The connected links does not necessarily focus on nearest hospital. The connection is created to maximize the resources distribution based on the given risk (or, disaster locations in this case) while still limited by the travel time constraint W.

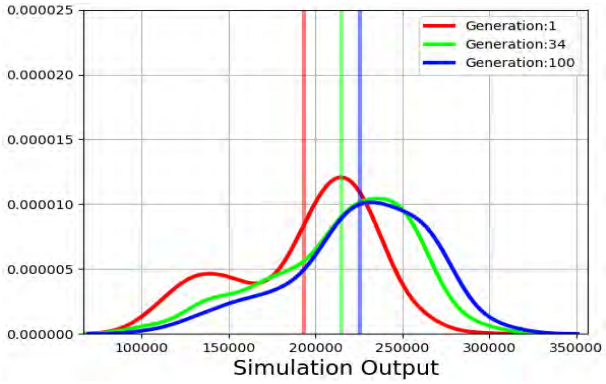


Fig. 1. Density plot of the simulation results of 500 scenarios for 1st, 34th, and 100th generation hospital networks. The vertical line indicates the network/chromosome performance.

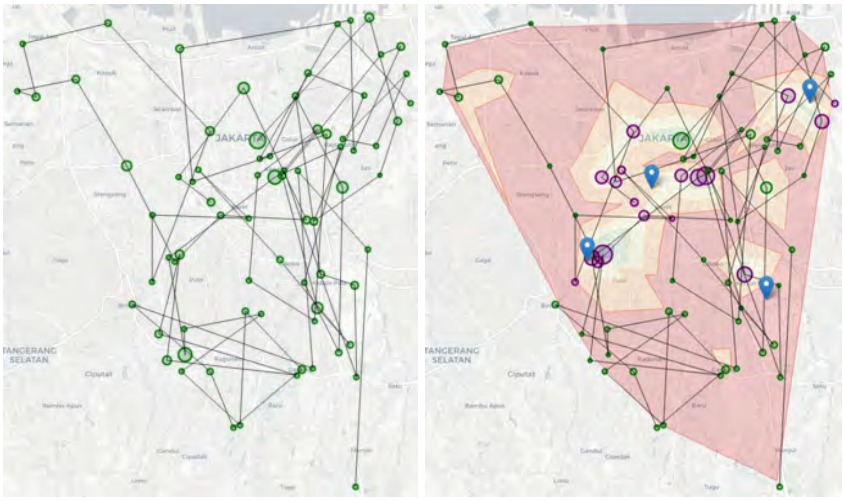


Fig. 2. (Left) The optimized hospital networks. (Right) A random disruption scenario were generated (blue marker). Purple and green nodes are the hospitals within and outside the disaster vicinity, respectively. Node size is based on the resources value. The contour indicates the level of resources in the area after the transfer of medical staff.

References

1. Amir, S. and Kant, V.: Sociotechnical resilience: a preliminary concept. Risk Analysis, 38(1), pp.8-16 (2018)
2. Barbera, JA., Macintyre, AG.: Medical Surge Capacity and Capability. U.S. Department of Health and Human Services (Sep 2007)

A Determinant Criterion for Stability Analysis of Complex Systems

Chandrakala Meena[1], Haber Simcha[1], Chittaranjan Hens[2], and Baruch Barzel[1]

[1] Mathematics Department, Bar-Ilan University, Ramat-Gan, 5290002 Israel,
meenachandrakala@gmail.com,
[2] Indian Statistical Institute, Kolkata, West Bengal 700108, India

1 Introduction

Prediction of the dynamic stability of a complex system is a challenging problem these days. Spectral graph theory is only good for a linear dynamics, while complex systems incorporates nonlinear interactions. So, current stability analyses fails to account for the actual stability of real complex networked systems. In our work, using our *Dynamic exponents* we retrieve the structure of Eq. (1) is real stability matrix, exposing the system's true stability profile. The analytical prediction of topological and dynamic parameters that govern the system's large scale behaviour, namely *Dynamic exponents*, provide the currently lacking bridge between the two mapping known *Topological elements* into desired *Observable* outcomes. Here our main objective to understand the *Dynamic exponents* of stable vs. unstable states. To determine these *exponents* we use a framework of complex system that incorporates two layers (i) *topology*, A_{ij}, which captures the geometry supporting the interacting elements and *Dynamics*, $\mathbf{M} = (M_0, M_1, M_2)$, capturing the interaction mechanisms between the nodes. Most broadly, these two layers translate to [1]

$$\frac{dx_i}{dt} = F_i(\mathbf{x}) = M_0\big(x_i(t)\big) + \sum_{j=1}^{N} A_{ij} M_1\big(x_i(t)\big) M_2\big(x_j(t)\big) \tag{1}$$

where $M_0(x)$ represents i's self dynamics, and the $M_1(x), M_2(x)$ capture the system's pairwise interaction mechanisms. Using a set of analytical arguments established in [1, 2] we show analytically that for systems of the form Eq.(1) the magnitude of Jacobian terms is not random, rather its term are determined by the interplay between Topology A_{ij} *and* Dynamics $\mathbf{M}$, via the scaling relations

$$J_{ij} = \left.\frac{\partial \dot{x}_i}{\partial x_j}\right|_{\vec{x}} \sim S_i^{\beta} S_j^{\gamma}, \quad J_{ii} \sim S_i^{\alpha} \tag{2}$$

where α, β, γ are *Dynamic exponents* of the system, analytically tractable from Eq. (1). Eq. (1) obtained in the asymptotic limit of $N \to \infty$, captures the impact of the degrees S_i and S_j on the relevant terms J_{ij}, while the degrees depend on A_{ij}. The system is asymptotically stable if real part of the largest eigenvalue λ_1 of Jacobian J_{ij}, is negative [3].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

Eq. (2) predicts that, despite using the same A_{ij}, each of these systems will have a fundamentally different Jacobian structure due to the nonlinear **M**, and its associated *Dynamic exponents* α, β, γ. To test this we numerically obtained J_{ij} (red circles) and confronted with Eq. (2)'s analytically predicted scaling (blue solid lines). The results in Fig. 1(e) - (i); fully corroborate our predictions, but, most importantly, they show that the real Jacobian matrices, as obtained from A_{ij} *and* the nonlinear **M**, are profoundly different than the commonly assumed random matrix ensembles. In fact, they are profoundly different from each other, due to the Dynamics, despite the fact that the underlying Topology is the same in all cases. As expected, the spectra $P(\lambda)$ of these Jacobians in Fig. 1 (j) - (l) takes diverse forms, a unique *fingerprint* of each network's dynamics.

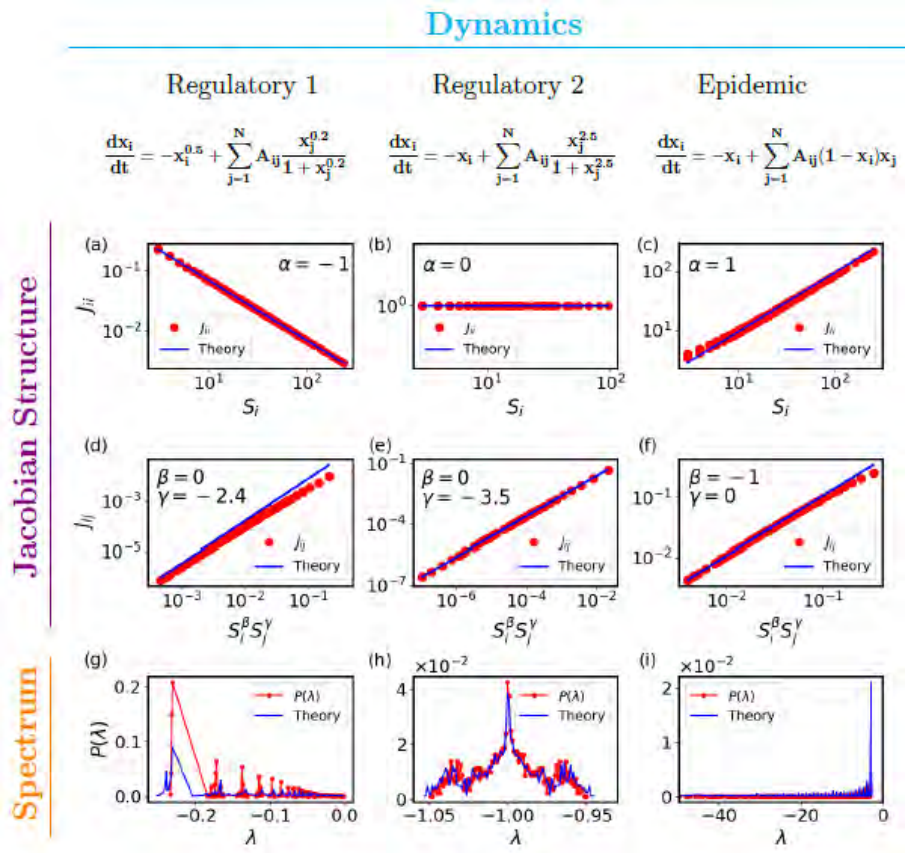


Fig. 1. Analysis of Jacobian matrices. Here we examine a scale-free A_{ij} [4] coupled with three different dynamics: (a) - (b) Gene regulatory dynamics [5] with different parameters and (c) susceptible-infected-susceptible [6] model for epidemic spreading in Eq. 1. (d) - (i) Numerically obtained J_{ij} (red circles) and confronted with analytically predicted scaling (blue solid lines). (j) - (l) show eigen spectrum $P(\lambda)$ of these Jacobians with *Dynamic exponents* α, β and γ.

The crucial point is that Eq. (2) introduces a link between the *structure* of J_{ij}, as determined from the Topology (A_{ij}, and hence S_i, S_j), and the *magnitude of its specific entries*, as affected by the Dynamics (α, β, γ). Hence *real* J_{ij}, *i.e.* ones derived from actual dynamics, are *not* extracted from the random matrix ensemble, but rather from a profoundly different and currently unexplored ensemble, in which structure and dynamics are deeply intertwined. Our scaling relation is mathematically solid because we have tested it for many other dynamical systems also for instance, population [7], mutualistic ecology [8] dynamics. Many research have been going on complexity-stability paradox from few decades. *Our piece of work shed a new light on this diversity-stability debate. Our results are uncovering the true statistics of complex systems that shape the structure of J_{ij}, and hence predicting the system's actual response to linear perturbations*.

Summary. We have uncovered a small set of *Dynamic exponents* that help us in translation of static structure into a dynamic stability matrix. With these exponents we found that similar networks may have very different dynamic spectra depending on their dynamics. These spectra are fully predicted by the small number of exponents, so we can predict precisely the stability of the system. These spectra are profoundly different from the currently explored random ensembles, thus offering a conceptually novel solution to the May's stability paradox [9] eluding decades old challenge. Thus our results in a richness of potential dynamic behaviors that is unaccounted for within the classic random matrix framework. Now we can also understand the organizing principles of stability and the crucial interplay between structure and dynamics.

References

1. Barzel, B. and Barabási, A.-L.: Universality in network dynamics. Nature physics 9(10), 673–681 (2013)
2. Barzel, B., Liu, Y-Y and Barabási, A.-L.:Constructing minimal models for complex system dynamics. Nature communications 6, 7186 (2015)
3. Coyte, K. Z, Schluter, J. and Foster, K. R: The ecology of the microbiome: networks, competition, and stability. Science 350(6261), 663–666, (2015)
4. Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512, (1999)
5. Alon, U.: An introduction to systems biology: design principles of biological circuits. Chapman and Hall/CRC, London, U.K. (2006) (1961)
6. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. and Vespignani, A.: Epidemic processes in complex networks. Rev. Mod. Phys. 87(3), 925–958, (2015)
7. Novozhilov, A. S., Karev, G. P. and Koonin, E. V.:Biological applications of the theory of birth-and-death processes. Briefings in bioinfo. 7(1), 70–85 (2006)
8. May, R. M.:Simple mathematical models with very complicated dynamics. Nature 261(5560), 459–467 (1976)
9. May, R. M.:Will a large complex system be stable?. Nature 238(5364), 413–414, (1972)

Multilayer networks meet databases: v/e-cubes as the building blocks of networks

Matteo Magnani

InfoLab, Dept. of Information Technology, Uppsala University, Sweden
matteo.magnani@it.uu.se

1 Introduction

Databases are fundamental components of typical data analysis systems and processes. While data analysis algorithms from statistics and machine learning are often defined on a given input, such as a data table, databases allow us to store the raw data, to enforce constraints guaranteeing some levels of data quality, and to efficiently manipulate (or query) the raw data to obtain the data table(s) needed for the analysis. This is particularly important when performing interactive and visual data analysis, where several data dimensions, data subsets and aggregations have to be added and removed dynamically to explore different views over the data.

In this work we present an extension of the multilayer network model [1] providing database functionality on top of the idea of using the concept of layer to unify several types of graph-based models. While the main contribution of this work is to provide data manipulation operators for multilayer networks, allowing us to dynamically create layers and tranform the data, our model also extends some features of the original multilayer model. First, not only the vertices but also the edges are allowed to be grouped into different layers, which makes the application of the model more intuitive. Second, attributes on vertices and edges as well as the ordering between layers are part of our model and can be directly manipulated e.g. using an ORDER BY operator, whereas they are left to custom software implementations in the original multilayer model. Similarly, in the original multilayer model there is no explicit concept of data constraint to specify the validity of a given network.

2 v-cubes and e-cubes as network model building blocks

If we use a (simple) attributed graph as an established data model for networks – what is known as a *property graph* in the database world, then we can represent data as a set of vertices and a set of edges associated to these vertices (V, E), in addition to attributes. In a multilayer attributed network, vertices (and also edges in our case) exist inside layers. As in the original multilayer model, layers are defined by aspects (called dimensions in the database world, in particular in multidimensional databases and data warehousing), resulting in vertex and edge cubes. Please notice that despite the similar names the cubes defined in this work are clearly distinct from the concept of graph cube presented in [3], where cube operations are based on aggregation (change of granularity), and edges are not directly manipulated.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

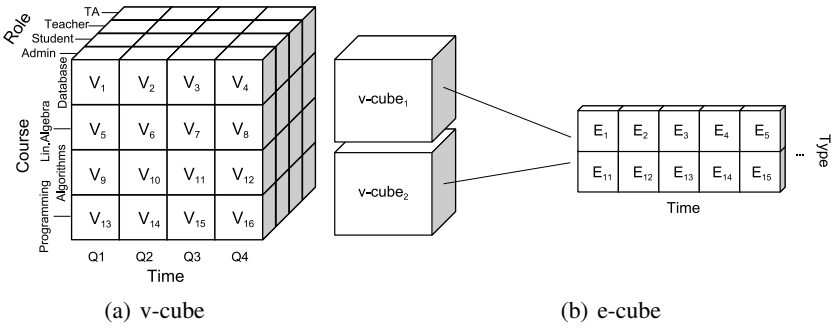


Fig. 1. A v-cube with three dimensions and an e-cube defined on two v-cubes

Fig. 1(a) shows an example of vertex cube over three aspects/dimensions, where each layer/cell contains a set of vertices and the same vertex can be present in multiple cells. Edge cubes are defined on pairs of vertex cubes (which can also be the same vertex cube, or slices of a larger vertex cube), as in Fig. 1(b). The two ends of the edges stored in an edge cube must belong to the two vertex cubes.

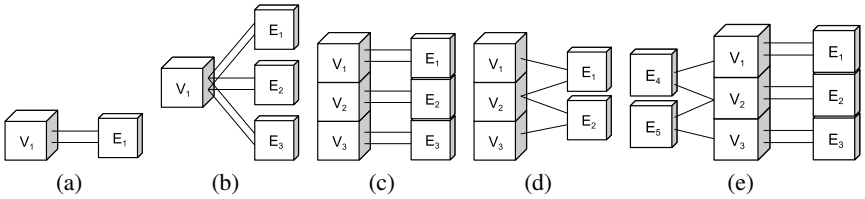


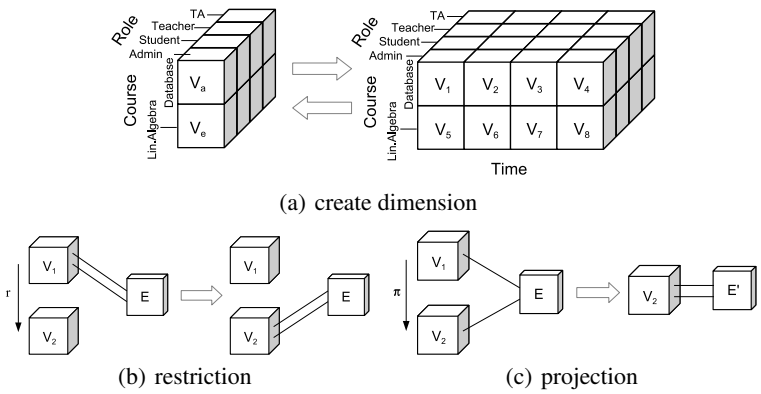
Fig. 2. Existing models represented as combinations of graph cubes: (a) Property graph, (b) Multi-relational, (c) Generalized multiplex, (d) Multi-mode and (e) General multilayer

Fig. 2 shows some examples of how popular network models can be expressed as combinations of v-cubes and e-cubes. The different configurations also define the domains of the edges, constraining the data that is valid for each configuration.

3 Operators

For both v-cubes and e-cubes we can re-use the same operations defined for data cubes in data warehousing, e.g., roll-ups, drill-downs, slicing, dicing, re-ordering and pivoting, not shown here for space reasons. In addition we can define network-specific operators, three of which are shown in Fig. 3. (a) We can create new dimensions using a function that given a vertex/edge indicates where in the new dimension the vertex/edge would be present. An example is a topic detection function that classifies social media posts into one or more topics. Another example is the dynamic creation or temporal slices. (b) When a v-cube has been manipulated (e.g., sliced), we can restrict e-cubes

defined on the original vertices to the new v-cube. (c) A third example are projection operators, expressing edges in two-mode networks as edges within one mode.



(a) create dimension

(b) restriction

(c) projection

Fig. 3. Some common operations on a (data/v-/e-) cube

4 An example application

After presenting the full data manipulation language, we will conclude the presentation by showing how to express a process to identify online conversations about specific topics in social media as a combination of cubes and cube operators, as done in [2]. This will be based on political Twitter communication data in Denmark and Sweden.

Acknowledgments This work was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 727040 (Virt-EU).

References

1. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer Networks. Journal of Complex Networks 2(3), 203–271 (2014)
2. Vega, D., Magnani, M.: Foundations of Temporal Text Networks. Applied Network Science 3(1), 25:1–25:26 (2018)
3. Zhao, P., Li, X., Xin, D., Han, J.: Graph cube: on warehousing and OLAP multidimensional networks. In: SIGMOD Conference. pp. 853–864. ACM (2011)

Part XVIII

Social Networks

Disasters and Polarization in Social Media

Güneş Ertan

Koç University, 34450, Istanbul, Turkey,
gunesertan@ku.edu.tr

1 Introduction

Despite increasing recognition of the relationship between social media and extreme events scholarly research in this area is still limited. Existing research mostly focuses on the function of social media in various phases of disaster management such as spreading disaster preparedness information, providing early warnings, assessment of disaster damage, facilitating response operations, providing actionable information to first responders, and enabling self organization among citizens and communities [1, 2]. In this paper we aim to make two novel contributions: (1) Demonstrate that social media during and after a disaster can also reveal fault lines of polarization in the society; and (2) provide an alternative approach to measure polarization based on co-hashtag networks during and after a disaster. Data for the study is based on tweets about the Soma mining disaster in Turkey. On May 13th 2014, 301 workers died at the Soma Holding coal mine in the Soma town of Manisa, Turkey. The incident is the deadliest mining as well as workplace accident in the history of the modern Turkey [3] . We use Twitter data accessed through Gnip, the subsidiary of the Twitter company. We analyze all tweets that include the word "soma" or hashtag "soma" following the first three days of the disaster. This time period starts with the first emergency call from the mine on May 13th and ends on May 17th, 3pm, the time that the last miners body was removed from the mine . The volume of Twitter activity is presented in Figure 1.

We conceptualize polarization as a social phenomenon characterized by presence of cohesive subgroups that have clashing views and positions along the lines of political parties and ideologies with small number of individuals or organizations as intermediaries between groups. Existing research shows strong presence of polarization in social media platforms such as Twitter [4]. There is no scientific consensus on how to best operationalize and measure polarization in social media. The most common approach to measure polarization in social media is to first estimate the group membership of each node in the network based on follower/friendship networks or text analytic techniques, and use community detection algorithms [5]. However application of techniques to non-English contexts as well as to multiparty systems is limited.

2 Results

For the analaysis we used hashtags that are tweeted at least 10 times.We constructed a hashtag network by considering a link between two or more hashtags if they are used in the same tweet. After removing the isolates, we utilized Louvain multi-level modularity optimization algorithm to identify community structures within the co-hashtag

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

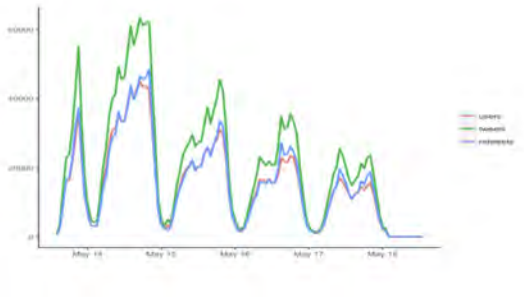


Fig. 1. Tweeter Activity on Soma Disaster May 13th-May 17th

Fig. 2. Hashtag Communities- Soma Disaster

network [6]. This algorithm generated 13 communities as displayed in Figure 2. The hashtags in the two largest communities convey substantially different meanings. The first group mostly consist of religious sentiments and prayers whereas the second group display anger, resentment, and calls for protests and boycotts around the country. Next we analyze individuals tweeting behavior in order to see to what extent same individuals tweet in different hashtag groups. Table 1 shows the Euclidean distance between top five hashtag communities based on users' number of tweets in each group. The dissimilarity between the group that is dominated by anger and call for protest hashtags, and the group in which religious sentiments and prayers are the majority is relatively high. This finding suggests that users tweeting in the religious sentiments group is not very likely to participate in the anger/protest group or vice-versa. These dissimilarities may also be indicative of polarization in the sense-making processes of Twitter users.

	Group A	Group B	Religious Sentiments/Prayers	Group C	Anger /Protest
Group A		1175	1373	1188	1276
Group B			1206	797	895
Religious Sentiments/Prayers				1171	1156
Group C					749
Anger/Protest					

Table 1. Euclidean Distance between Hashtag Communities

Summary. This paper shows social media behavior during and after a disaster may be indicative of political polarization. We also provide an alternative approach to measuring polarization based on co-hashtag networks. This approach may be useful for analysis of non-English social media text data for which text analytic tools are considerably limited.

References

1. Murthy, D., Gross, A.J.:Social media processes in disasters: Implications of emergent technology use.Social science research. 63, 356–370 (2017)
2. Imran, M., Castillo, C., Diaz, F., and Vieweg, S. Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR). 47(4), 67 (2015)
3. Erkan, B., Ertan, G., Yeo, J. and Comfort, L.K.: Risk, profit, or safety: Sociotechnical systems under stress. Safety science. 88, 199–210 (2016)
4. Tucker, J.A., Guess, A., Barber, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D. and Nyhan, B. Social media, political polarization, and political disinformation: A review of the scientific literature. March , 2018
5. Guerra, P.C., Meira Jr, W., Cardie, C. and Kleinberg, R. A measure of polarization on social media networks based on community boundaries. In Seventh International AAAI Conference on Weblogs and Social Media. June 2013
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E.:Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. 10, P10008 (2008)

Friendship Paradox and Hashtag Recommendation in Instagram

David Serafimov, Igor Mishkovski, Sasho Gramatikov, and Miroslav Mirchev

Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University in Skopje, North Macedonia

1 Introduction

People involved in everyday online social networking subjectively think that they have more friends than most of their friends. However, in 1991 it was discovered that "most people have fewer friends than their friends have, on average" [2]. This phenomenon, known as *friendship paradox*, does not only hold for the number of friends in social networks, but also for other properties, such as activity, virality happiness, number of partners, etc. The friendship paradox is both statistical and behavioral [3], and its existence is already shown in Twitter [4] and Facebook [5], but to the best of our knowledge it has not been analyzed in details for Instagram. Using a dataset extracted from Instagram [6], we investigate the existence of this phenomenon on different properties, such as likes, comments, posts and hashtags on the followers and followees Instagram sub-network. Using the same data set, we additionally represent hashtags as vectors/embeddings in order to provide hashtags recommendation, which to the best of our knowledge have not be done elsewhere. In our study, we rely solely on the available hashtags, unlike several approaches in the literature where authors use multidimensional features like images [7] and text [8]. This unique approach allows us to easily adapt them for data sets where such multidimensional features are not available. This approach can be applied for exploration, control and prevention of spreading trends, represented as hashtags, across a network [9]. The results from this work have been already accepted for publication in [1], where the reader can find more details, while here we present the main findings.

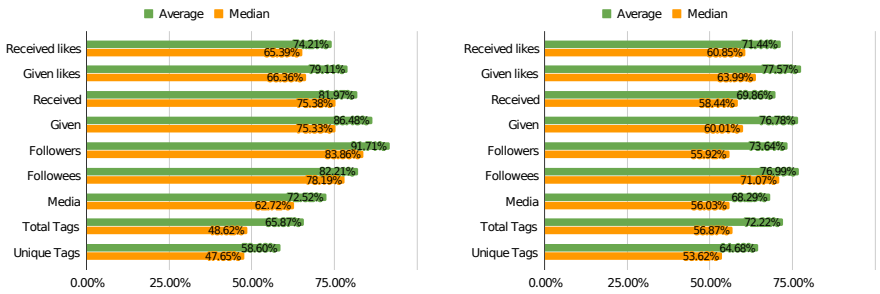
2 Friendship Paradox in Instagram

There are plenty of directed activities in the Instagram network, such as follow users, allow to be followed, post images, videos, likes, comments, etc. In this sense, we rephrase the friendship paradox as: i) Our *followees* have or do something more than us on average, and ii) Our *followers* have or do something more than us on average. In addition, we consider two types of paradoxes: *weak* paradox - the user's activities are compared to his neighbor's average (related to statistical aspect) and *strong* paradox - the user's activities are compared to his neighbor's median (related to behavioral aspect).

From Fig. 1 we can observe that the weak friendship paradox represented as a percentage, exists for all of the analyzed activities when applied both to the user's followees (see Fig. 1(a)) and followers (see Fig. 1(b)). On the other hand, its strong variant does not hold only for the total and unique hashtags for the user's followees (see Fig. 1(a)).

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



(a) relating to the followees (b) relating to the followers

Fig. 1: Percentage of users for which the friendship paradox applies for various properties relating to: (a) the followees and (b) the followers.

By comparing Fig. 1(a) and Fig. 1(b) we see that, except for the number of hashtags, the friendship paradox is stronger for the user's followees than his followers for all social activities. This means that, unlike our followers which use numerous and diverse hashtags, our folowees prefer the use of fewer, but more specific hashtags.

3 Hashtag Analytics and Recommendation

The usage of Natural Language Processing (NLP) on hashtags is helpful in Named Entity Recognition (NER), sentiment analysis and other classification tasks in Instagram. Several approaches solve this tasks on Instagram using additional information like images [7] or text [8]. Our approach is based solely on hashtags embedded in a multidimensional space using the Word2Vec method [10], making it more general.

In order to represent the hashtags in a vector space, we use neural network trained with the hasthags that appear in a single post. The network has a hidden layer with size defined by the desired dimension of the representing vector, whereas the input and output layer are defined depending on the specific model we use. The "bag of hashtags" (BoH) method has as many inputs as hasthags in the post reduced by one, and one output which is the excluded hashtag, while the "hashtag pairs" (HP) method has one input hashtag and one output which is one of the surrounding hashtags. After preprocessing the Instagram dataset (hashtags with less than 3% occurrences were filtered) and splitting it into 90% training and 10% test sets, we train the two models with 64 and 128 nodes in the hidden layer over 50 epochs and compare the results with a baseline statistical (BS) model that gives recommendations according to previous occurrences of different hashtags. After training, the coefficients obtained for each hashtag that refer to the hidden layer form the vector representative of that hashtag. Our evaluation employs the Recall at K (R@K) metric to measure the models' quality (which is the average number of relevant hashtags recommended in the top K.). From Table 1 we can conclude that the HP model outperforms the BoH model. Using the resulting vectors, which represent hashtags, we can further calculate hashtag similarity, search similar hashtags,

Table 1: The recall at K (R@K) metric

Model	R@1	R@2	R@3	R@5	R@10
BS	0.0201	0.0366	0.0480	0.0703	0.1184
64D BoH	0.0617	0.0865	0.1035	0.1274	0.1661
64D HP	0.0779	0.1157	0.1435	**0.1836**	**0.2414**
128D BoH	0.0339	0.0477	0.0572	0.0713	0.0956
128D HP	**0.0824**	**0.1189**	**0.1445**	0.1801	0.2307

cluster the hashtags and get hashtag topic extraction on Instagram. We can also perform arithmetic operations, such as: #helloween − #pumkin + #christmas = #christmastree and #sweden − #stockholm + #turkey = #istanbul

Summary. In this work we prove that strong and weak friendship paradox exist for the number of followers, likes, posts, hashtags and comments, both regarding the followers and the followees. Moreover, the friendship paradox is more obvious for user's followees rather than for his followers, excluding the number of hashtags. In addition, we introduced a general method for obtaining high-quality hashtag representations in multidimensional space and tested the obtained hashtag embedding on hasthag recommendation. Compared to the baseline model our method achieved better results.

References

1. D. Serafimov, M. Mirchev, and I.Mishkovski: Friendship paradox and hashtag embedding in the Instagram social network. 11th ICT Innovations conference (2019) (accepted)
2. L. Feld. Scott: Why your friends have more friends than you do. American Journal of Sociology, 96(6):1464-1477 (1991)
3. K. Farshad, N. O. Hodas, and K. Lerman.: Network weirdness: Exploring the origins of network paradoxes. 8th Int. AAAI Conf. on Weblogs and Social Media (2014)
4. N. O Hodas, F. Kooti, and K. Lerman.: Friendship paradox redux: Your friends are more interesting than you. In 7nth Int. AAAI Conf.e on Weblogs and Social Media (2013)
5. K. N. Hampton, L. S. Goulet, C.Marlow, and L. Rainie.: Why most Facebook users get more than they give. Pew Internet & American Life Project, 3:1-40 (2012)
6. E. Ferrara, R. Interdonato, and A. Tagarelli.: Online popularity and topical interests through the lens of instagram. In Proc. of the 25th ACM conf. on Hypertext and social media, pages 24-34 (2014)
7. A. Veit, M.n Nickel, S. Belongie, and L. van der Maaten.: Separating self-expression and visual content in hashtag supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5919-5927 (2018)
8. J. Weston, S. Chopra, and K. Adams.: # tagspace: Semantic embeddings from hashtags. In Proc. of the 2014 conference on EMNLP, pages 1822-1827 (2014)
9. L. Zhang, J. Zhao, and K. Xu.: Who creates trends in online social media: The crowd or opinion leaders? Journal of Computer-Mediated Communication, 21(1):1-16 (2015)
10. T. Mikolov, K. Chen, G. Corrado, and J. rey Dean.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

The Network Structure of Freelance Journalism

Nick Hagar and Emőke-Ágnes Horvát

Northwestern University, Evanston IL 60208, USA
nhagar@u.northwestern.edu and a-horvat@northwestern.edu

1 Introduction

Over the first two months of 2019, at least 2,200 people lost their jobs in the news media industry across the U.S. [4]. As full-time media positions disappear across the globe, companies are turning to freelance labor to fill the gap, changing the media labor landscape in the process. For example, Vox Media has expanded its job listings for contract and freelance positions without movement for full-time jobs [2]. As the industry faces ongoing financial strains, the security of labor available to media workers is shifting.

Those shifts raise broader questions around the structure of unstable labor markets. Almost all freelance journalists work on assignments for multiple publications, meaning they constantly move among organizations [8]. As contract work becomes more prevalent across other industries—as in areas like transportation and food delivery—the free flow of workers among companies becomes increasingly important.

Within news media, the flow of workers also influences the dissemination of information and viewpoints. Past work has shown how consumption and distribution patterns isolate news along political lines in the U.S., particularly in right-leaning media [1]. If freelance journalists cannot move freely among publications across the political spectrum, news production could also demonstrate this polarization, limiting which perspectives reach certain audiences.

This problem is well suited for a complex networks perspective. By analyzing the connections formed between organizations when they share workers, we can begin to understand the structural characteristics of movement among companies in an industry that is fundamental to our democracies. Journalism is an excellent case for studying this question because journalists' articles written in various publications enable tracking their professional trajectories at a large scale.

This study examines the structure of digital journalism's labor market through the lens of 401 writers' publishing histories over 3 years (June 2014 through July 2017, most stories published in 2016-2017). Using a comprehensive sample of 6,567 news stories across 14 major outlets such as The New York Times and Washington Post [9], we construct a weighted network of digital publication connections based on how many writers they shared. From this network, we first demonstrate that writers' publishing trajectories tend to follow common patterns across news outlets. Then, we show that those patterns most closely align with the political leanings of the outlets in our sample, rather than more traditional delineations like outlet medium or audience characteristics.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

2 Results

To test for the presence of structural patterns of movement among outlets, we identify significant relationships between outlets based on the writers they shared [5]. We compared the one-mode projection of the journalist–outlet bipartite network, weighted by the number of shared journalists between each pair of outlets, to degree-preserving permutations [3], identifying those pairs that shared more writers than expected by chance based on the bipartite network's structure.

Figure 1 shows the significant relationships among outlets ($p < 0.05$). These connections first demonstrate that writers follow patterns in where they publish. This indicates a structural trajectory that can determine the broad contours of a freelance journalist's career. Second, this analysis identifies two distinct clusters of publications: One low-density cluster that comprises nine publications and a clique of four outlets.

At first glance, these patterns don't seem to align with common industry segmentations. Traditionally, work that examines multiple outlets has divided them by their publishing medium—comparing print newspapers to news websites, for example, or radio to television [7]. In this network, though, medium does not seem to align with the clusters we observe: Magazine contributors also write pieces for radio and digital-native sites, for example (e.g., journalists shared by The Atlantic, NPR, and Vox).

To validate this observation, we categorize each outlet using a variety of metadata: political leanings from Allsides [6], the city of its headquarters, audience size, age, and income data from comScore (data collected for April 2019 and divided in quartiles), and a broad manual coding of the outlet's traditionally dominant medium (print, digital, radio, television, or wire service). We then measure how frequently the freelancers in our sample transition between categories for each classification, as a measure for how well each describes the observed clusters.

Figure 2 shows the mean number of transitions per writer for each classification, along with a 95% confidence interval. We find that the political leanings classification has the lowest number of transitions per writer on average, meaning it best captures the clustering seen in our network. Indeed, if we map the leanings of each publication to the network in Figure 1, we see a dense cluster of right-leaning publications and a loose grouping of left- and -center leaning ones. This finding suggests that the political leaning of a news outlet, especially for right-leaning publications, is a key structural factor in where freelance journalists will publish during their careers.

These results start to reveal a structural component to whether or not freelancers find success in various outlets. Beyond the strength of the pitch and the track record of the individual journalist, publications may look for specific peer institutions when evaluating freelance labor. Those heuristics have the potential to create inequality among freelancers. The results also raise potential concerns around who is producing right-leaning media in the U.S. If the freelancers who write for mainstream news sites don't also write for conservative media, and vice versa, that means on the one hand that right-leaning sites are not drawing on the industry's main freelance talent pool. On the other hand, if more moderate writers are not contributing to those sites, or are stigmatized by the broader industry for doing so, media production will see continued polarization. Our study also indicates that organizational characteristics contribute to broader industrial structures, which in turn shape and constrain the movement of freelancers and contrac-

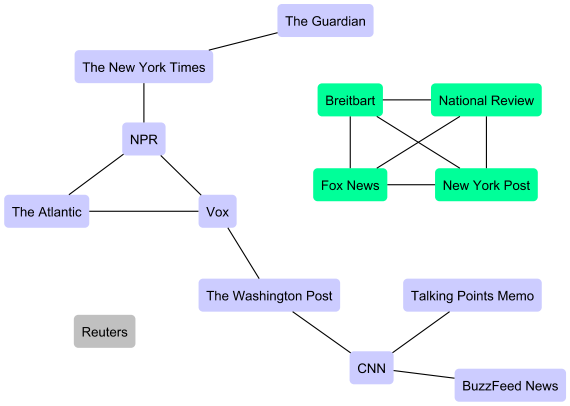


Fig. 1. Statistically significant connections observed among the 14 publications in our sample ($p < 0.05$), colored to indicate distinct clusters.

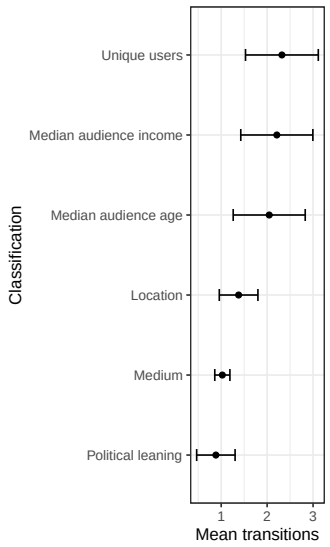


Fig. 2. Mean transitions per writer for each classification.

tors. Understanding the structure of an industry's labor movement network is therefore crucial for assessing what opportunities are available to which workers.

References

1. Benkler, Y., Faris, R., Roberts, H., 2018. Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press, New York, NY.
2. Fruhlinger, J., 2019. Vox Media Ramps Up Hiring, But Mostly for Contractors and Freelancers. Thinknum.
3. Gionis, A., et al., 2007. Assessing Data Mining Results Via Swap Randomization. ACM Trans. Knowl. Discov. Data 1, 14.
4. Goggin, B., 2019. More Than 2,200 People Lost Their Jobs in a Media Landslide So Far This Year. Business Insider.
5. Horvát, E.-Á., Zhang, J.D., Uhlmann, S., Sahin, Ö., Zweig, K.A., 2013. A Network-Based Method to Assess the Statistical Significance of Mild Co-Regulation Effects. PLoS ONE 8(9), e73413.
6. How AllSides Rates Media Bias: Our Methods, 2016. AllSides. URL https://www.allsides.com/media-bias/media-bias-rating-methods
7. Maier, S., 2010. All the News Fit to Post? Comparing News Content on the Web to Newspapers, Television, and Radio. Journalism & Mass Communication Quarterly 87, 548–562.
8. Rosenkranz, T., 2018. From Contract to Speculation: New Relations of Work and Production in Freelance Travel Journalism. Work, Employment and Society 33, 613–630.
9. Thompson, A., 2017. All the News: 143,000 Articles from 15 American Publications. URL https://www.kaggle.com/snapcrack/all-the-news

Does diversity kill OSNs?

László Lőrincz[1], Júlia Koltai[2,3], Johannes Wachs[4], and Károly Takács[2,5]

[1] Hungarian Academy of Sciences, Centre for Economic and Regional Studies, Institute of Economics, Tóth Kálmán utca 4., H-1097 Budapest, Hungary, and Corvinus University of Budapest, Fővám tér 8, H-1093 Budapest, Hungary
[2] Hungarian Academy of Sciences, Centre for Social Sciences, Computational Social Science – Research Center for Educational and Network Studies (CSS – RECENS), Tóth Kálmán utca 4., H-1097 Budapest, Hungary
koltai.julia@tk.mta.hu
[3] Eötvös Loránd University of Sciences, Faculty of Social Sciences, Pázmány Péter sétány 1/A., H-1117 Budapest, Hungary
[4] Chair of Computational Social Sciences and Humanities, RWTH Aachen, Theaterplatz 14 52062 Aachen, Germany
[5] The Institute for Analytical Sociology (IAS), Linköping University, S 601 74 Norrköping, Sweden

Keywords: Online Social Networks, context collapse, social capital, diffusion of innovations, network resilience, network cascade

1 Introduction

In the offline world, one has contact with diverse groups of people. The composition of family members, old friends from school, colleagues, neighbours, and other acquaintances can be very different set of people, also with regard to their attitudes and political preferences. Some cultural codes and meanings vary between these groups: opinions and norms that are acceptable in one group can be condemned in another. These differences can happen in the daily life (e.g., when husband and wife work together), however, members of the different groups do not meet that often; such as the role of ego is different in the different groups, and the acting out of these roles usually meets the expectations of the group. In contrast, the online sphere creates large space for opening boundaries between the diverse groups of the ego. If someone posts content to social media, all contacts from the diverse set of people can see and react to it. Thus, conflicting expectations arise. In the words of Marwick and boyd [NM&S, 2011], networked audiences we encounter online moderate our behavior and activity in different ways than face-to-face audiences.

The growing popularity of online social networks (OSNs) therefore creates more and more opportunity for context collapse. In this study, we examine the role of context collapse on the formulation of online communities, namely that how the existence of diverse groups affects if someone stays in or leave an online social network. In our planned project, we will analyse the role of context collapse in the abandonment of OSNs using the database of iWIW (international who is who, originally WiW). Earlier results explain and explore the mechanisms behind the leaving and collapse of the

online social network, iWIW from different points of views, however, the general role of context collapse has not been studied yet.

2 Data

iWiW was founded in 2002 and was one of the first online social networks in the world. At its peak, with its more than 3.5 million users, it was the biggest Hungarian OSN in a country with a population of 10 million, where this meant two-third of all Internet users. After 2010, with the appearance of Facebook iWiW started to lose its popularity and after the unstoppable decline in user activity, iWiW was finally shut down in 2014. While users of Facebook can also experience context collapse, we argue that it initially offered iWiW an escape from an increasingly heterogeneous environment. For some time early adopters of Facebook in Hungary were likely to encounter a more exclusive group of users on the new platform.

3 Methods

In our planned analysis we will examine the role of context collapse in the abandonment of the OSN. The dependent variable of the analysis is a binary variable that signs if in the given time-point the user is active or not active in the OSN. Based on our database, we can measure inactivity with two ways. One opportunity is the date of last login, at each user; the other option is the last date when the user added a new connection (friend).

The main independent variables of the analysis are the ones, which measure context collapse. For measuring context collapse, first we have to detect different groups of contacts at the ego-network of each user and then examine the homogeneity or heterogeneity of these groups from different point of views. For the detection of these groups of contacts we use Louvain community detection algorithm. This method uses modularity optimization and with its application, it is possible to typify the different groups of contacts for all users. After the detection of these groups, their comparison from different perspectives can be done at each user. These perspectives are limited by the database we use, but age, education and geographical location are available. Thus, the diversity of these contact groups is detectable from the perspective of education, location and age; and we are also able to calculate the distance between the user and the average value of the contact groups. The measurement of the diversification of these contact groups can be made by different similarity measures, which represent the appearance or the lack of context collapse at each user. These measures are the ones, which will be be included in the analysis as the main independent variables.

The other important factor in the analysis of the effect of content collapse on leaving the OSN is time. As the composition of the network is changing over time, we can treat our data as a dynamic structure. We take time into account on the level of years from 2007 to 2012. The unit of analysis will thus be person-years and we will include both fixed (e.g. year of birth) and random (e.g. context collapse) effect variables in the regression analysis, where we will treat the above described dataset as a panel data.

Over these variables, which measure context collapse, both fixed and random effect control variables will be included in the analysis, like gender, age of birth or number of contacts. Controlling for the latest mentioned variable is especially important, as the number of contact groups and thus the volume of context collapse strongly depend on the size of someone's network.

4 Expected results

We examine, how these groups of contacts facilitate individuals' decision on leaving the OSN, or which combinations of these circles contribute to individuals abandoning behaviour. Our results will provide contribution for the consequences of context collapse, which can add further results for the understanding of filter bubbles and echo-chambers.

ScamCoins, S*** Posters, and the Search for the Next Bitcoin^{TM}: Collective Sensemaking in Cryptocurrency Discussions

Eaman Jahani[1], Peter Krafft[1], Yoshihiko Suhara[1], Esteban Moro[1,2], and Alex 'Sandy' Pentland[1]

[1] MIT, Cambrdidge, MA, 02139, USA,
eaman,pkfrafft,suhara,emoro,pentland@mit.edu,
[2] GISC & Department of Mathematics, Universidad Carlos III de Madrid, Spain

1 Introduction

In the last years, cryptocurrencies have attracted massive attention from investors, institutions, policy-makers and the general audience. The public notoriety of Bitcoin, together with its sizable price increase, led to an explosion of attempts to create the *next Bitcoin*. Thus, a number of cryptocurrencies, often referred to as altcoins, and a vibrant set of exchanges have emerged particularly due to the extremely low cost and effort required to create or mutate a new coin, with some being minimal changes to parameters and branding of a pre-existing codebase. While many of these altcoins did not offer any new technological advancement, there have been some successful attempts in creating new cryptocurrencies that offered either significant technical innovation over the existing technology (e.g., Proof-of-stake in Peercoin) or introduced a wholly new idea (e.g. Turing Complete as in Ethereum) [3]. Given the abundance of new coins being created on a daily basis, it is natural to ask how well do traders detect cryptocurrencies that offer genuine technological innovation and are likely to succeed? A related question is whether the cryptocurrency community is attempting to collectively analyze and make sense of this large array of altcoins or is it simply engaged in hype-based speculation?

In this work [2], we use an empirical approach to assess whether and when the discussions of cryptocurrencies are truth-seeking or hype-based. We rely on a novel data set that combines measures of the main online forum discussion around cryptocurrencies with their price and volume history in exchange markets. Leveraging the literature on finance, we assume price represents the perceived fundamental value of a coin and treat its volatility as an indicator of information uncertainty around the technological innovation of the cryptocurrency. Similarly, drawing upon collective intelligence literature and using three measures of experience (seniority), information diversity (degree in the thread discussion network) and community engagement (equal participation by all community members), we quantify the extent to which the community discussion exhibits characteristics of collective sensemaking.

2 Results

Our results indicate a negative correlation between the quality of discussion measured in terms of collective sensemaking and price volatility of the coin suggesting that for

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

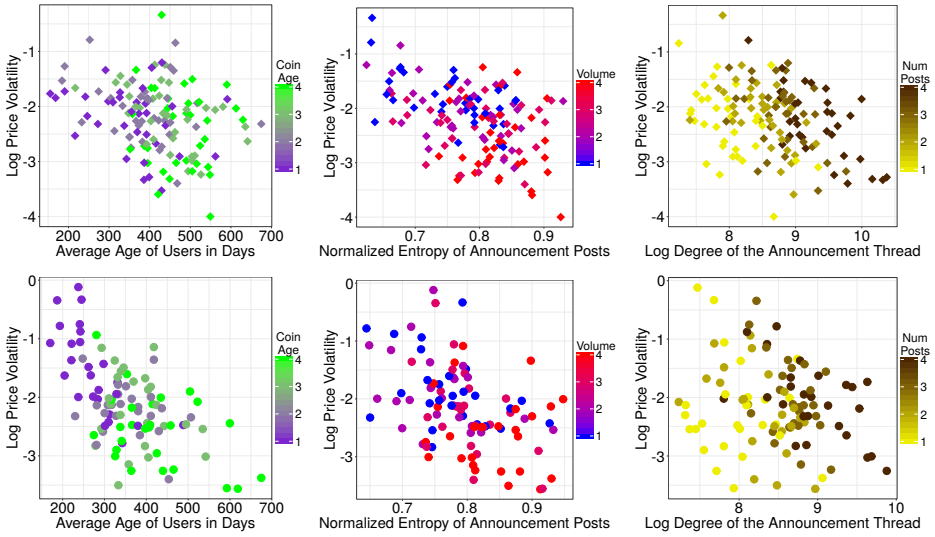


Fig. 1. The price volatility of the coins over two separate 100 day periods ending in November 2016 (top row) and January 2016 (bottom row) versus discussion variables: average age of the users in the discussion (left column), normalized entropy of number of posts made by each user (middle column) and the log degree of the discussion in the thread network (right column). All three discussion variables have a negative correlation with price volatility, a measure of information uncertainty.

"more serious" coins discussion is more likely to serve a truth-seeking role. Figure 1 shows that coins with more information available have equal participation by experienced contributors to the discussion (higher entropy of announcement posts made by various users) and more diverse opinions measured in terms of access to other information sources (Degree of coin announcement page). In contrast, coins with high information uncertainty tend to be discussed by less experienced and more narrowly focused users. We replicate the same results using an objective measure of technicality as a second operationalization of information uncertainty around the crypto coin. Table 1 shows the result of multiple one-sided two-sample t-tests for the hypothesis that the technical coins with more objective information available exhibit larger entropy, degree, user age and less price volatility than non-technical coins. We observe that the difference in means of all discussion quality variables is positive indicating that more technical coins have more substantial discussion.

The content analysis of the forum also reveals that the discussion of more innovative coins is more focused on the design and technical aspects. These results are consistent with qualitative findings of [1] and suggest that there are people in the cryptocurrency community who are mainly driven by market hype and view cryptocurrency as an investment, while others are dedicated to the technological advancement of the cryptocurrency ecosystem and view Bitcoin and its variants as a legitimate currency.

Summary. The public notoriety of Bitcoin led to creation of numerious cryptocurrencies, often referred to as altcoins. While many of these altcoins did not offer any new

Table 1. The one-sided t-test for the difference of means between technical and non-technical coins. The first row represents the observed empirical difference of means: $\overline{X}_{technical} - \overline{X}_{nontechnical}$. The second row shows the p-value of the Null hypothesis for the test H_0: *Non-Technical mean* $>$ *Technical mean* for each variable except volatility. In the case of volatility, the direction of the test is reversed and rejecting the Null for volatility indicates that volatility and technicality as operationalizations of information uncertainty are related as non-technical coins have higher volatility (uncertainty) than technical coins. Third row shows the 97.5% confidence interval for difference of means: $E[X_{technical}] - E[X_{nontechnical}]$. Results also indicate that coins with higher volume tend to be more technical and there is no significant relationship between coin age, total discussion acitivity and the coin technicality.

	Price Volatility	Entropy of Posts	Log Thread Degree	User Ages	Log Daily Volume	Coin Age	Log Total Posts
Means Diff.	-0.986	0.038	0.557	54.153	3.037	12.192	0.528
H_0 P-Value	0	0.0185	0.00296	0.00704	6.09×10^{-5}	0.453	0.102
97.5% CI	$(-\infty, -0.71)$	$(0.002, \infty)$	$(0.173, \infty)$	$(11.4, \infty)$	$(2.5, \infty)$	$(-194.4, \infty)$	$(-0.303, \infty)$

technological advancement, there have been some successful attempts in creating new cryptocurrencies that offered either significant technical innovation over the existing technology (e.g., Proof-of-stake in Peercoin) or introduced a wholly new idea (e.g. Turing Complete as in Ethereum). Given the abundance of new coins being created on a daily basis, it is natural to ask whehter the cryptocurrency community is attempting to collectively analyze and make sense of this large array of altcoins or is it simply engaged in hype-based speculation?

Our results suggest that there is at least a subgroup of online enthusiasts who are pursuing new technical coins and actively participate in their discussions if there is enough public information available. Our findings highlight the varied roles of discussion in the cryptocurrency ecosystem and suggest that discussion of serious coins may be oriented towards earnest, perhaps more accurate, attempts at discovering which coins are likely to succeed. In other words, as there is less uncertainty about the coin's technical merits, the discussion tends to become more truth-seeking. Finally, we hypothesize that the same discussion patterns may also be present in other forms of social media. In order to distinguish between hype, fake news, and similar noise, one can look at the character of the discussion surrounding the news item, and promote those that exhibit characteristics of collective intelligence.

References

1. Gao, X., Clark, G.D., Lindqvist, J.: Of two minds, multiple addresses, and one ledger: Characterizing opinions, knowledge, and perceptions of bitcoin across users and non-users. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 1656–1668. CHI '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2858036.2858049
2. Jahani, E., Krafft, P.M., Suhara, Y., Moro, E., Pentland, A.S.: Scamcoins, s*** posters, and the search for the next bitcointm: Collective sensemaking in cryptocurrency discussions. Proc. ACM Hum.-Comput. Interact. 2(CSCW), 79:1–79:28 (Nov 2018), http://doi.acm.org/10.1145/3274348

3. Ong, B., Lee, T.M., Li, G., Chuen, D.L.K.: Evaluating the potential of alternative cryptocurrencies. In: Handbook of digital currency. Elsevier (2015)

Statistical models of social interaction

Samuel Martin-Gutierrez[1], Juan C. Losada[1], and Rosa M. Benito[1]

Grupo de Sistemas Complejos, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Av. Puerta de Hierro, 2, 28040 Madrid, Spain.
https://orcid.org/0000-0002-5685-7834
https://orcid.org/0000-0002-4373-603X
https://orcid.org/0000-0003-3949-8232

Humans are organized in social systems, which implies that our actions as individuals hold the potential to trigger spontaneous reactions in our peers, leading to complex dynamics. In order to understand human collective behavior, it is necessary to find the laws that relate the individual actions to the collective response of social systems.

This topic has received considerable attention and has been approached from several perspectives [2, 1, 4, 5]. Our goal in this work is to unveil the mechanisms that drive the collective response of social systems to individual actions by developing models that describe the relationship between the intensity of the individual stimulus and the size of the collective response.

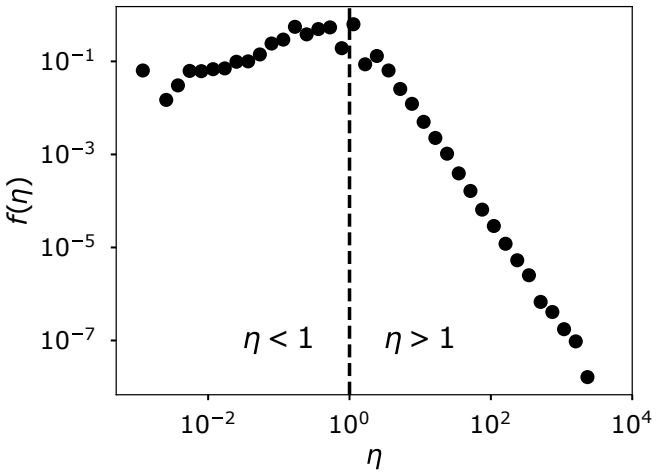


Fig. 1. Empirical efficiency distribution corresponding to a Twitter conversation.

We use the number of actions performed by an actor (an agent or individual) embedded in a social system; that is, her activity (A), as a proxy for the intensity of the individual stimulus. Likewise, we choose the number of reactions that are triggered by an actor in her peers, or response (R), as a proxy for the size of the collective response. To relate these two magnitudes we have generalized the efficiency metric, introduced by Morales et al. [3] in the context of Twitter, to other social systems, being the actor efficiency defined as $\eta = \frac{R}{A}$.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

We focus in studying parsimonious models to explain the empirical distribution of efficiency (see Fig. 1) for different social systems. To this end, we have developed three domain-independent statistical models that provide a description of how a social system reacts to the actions of its components. Each of them is based on minimal sets of assumptions that define different levels of dependence between R and A.

The models that we have developed are the Independent Variables model (InV), the Identical Actors model (IdA) and the Distinguishable Actors model (DiA). In the InV model the response of the system is independent with respect to the activity of the individual. In the IdA model, the response of the system depends on the activity of the individual, but the system is agnostic with respect to the individual that stimulates it. Finally, in the DiA model the response is determined both by the specific actor that performs the actions and by her activity.

We have applied these models to three social systems of different nature: Twitter conversations, the scientific citations network and the Wikipedia collaboration environment. The independent variables model captures the universal structure of the distribution of efficiency and explains its independence with respect to changes in the activity distribution, both empirical results found in previous works [3]. Additionally, it reproduces the efficiency distribution for the scientific citations network. The identical actors model improves the previous one by naturally inducing correlations between A and R that are comparable to those found in the data and reproduces the right tail of the efficiency distribution for the Twitter and Wikipedia datasets. Finally, the distinguishable actors model accurately fits the data for the whole range of efficiency.

References

1. Domenico, M.D., Altmann, E.G.: Unraveling the origin of social bursts in collective attention p. Preprint (2019), https://arxiv.org/abs/1903.06588
2. Juul, J.S., Porter, M.A.: Hipsters on networks: How a minority group of individuals can lead to an antiestablishment majority. Phys. Rev. E 99, 022313 (Feb 2019), https://link.aps.org/doi/10.1103/PhysRevE.99.022313
3. Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M.: Efficiency of human activity on information spreading on twitter. Social Networks 39, 1–11 (2014)
4. Muchnik, L., Pei, S., Parra, L.C., Reis, S.D.S., Andrade Jr, J.S., Havlin, S., Makse, H.A.: Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Scientific Reports 3, 1783 EP – (May 2013), https://doi.org/10.1038/srep01783, article
5. Rybski, D., Buldyrev, S.V., Havlin, S., Liljeros, F., Makse, H.A.: Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences 106(31), 12640–12645 (2009), https://www.pnas.org/content/106/31/12640

Opinion Polarization during dichotomous Twitter conversations

J.C. Losada[1], G. Olivares[1,2], J. Martnez-Atienza[1], S. Martn-Gutirrez[1], J.P. Crdenas[2], J. Borondo[1,3], and R.M. Benito[1]

[1] Grupo de Sistemas Complejos. Universidad Politcnica de Madrid. 28040 Madrid, Spain
[2] Net-Works, Angamos 451, Reaca, Via del Mar, Chile
[3] AGrowingData, Edif. PIF 1st f, Of. 5. Muelle de Poniente s/n, 04002 Almera, Spain

1 Introduction

Political polarization is a social phenomenon that has several consequences in peoples lives and whose nature is not completely understood. We say that a population is perfectly polarized when divided in two groups of the same size and opposite opinions. In this work, we have studied the polarization phenomena around Twitter conversations concerning different topics with clearly opposed opinions: electoral process with two candidates and social unrest. In each of the conversations, we have found a bipolar opinion distribution and a high value of the polarization index. In particular, we will present results of following Twitter conversations:

- The second round of the 2017 Chilean elections [1], where voters had to choose between the final two candidates.
- The Catalans independence process, where there are two clearly opposed positions (in favor and against independence)

To this end, we have built retweet networks and applied the model to estimate opinions proposed in [2], in which a minority of influential individuals propagate their opinions through the social network. The model takes into account two types of users, elites and listeners. In this model elite users have a fixed opinion that remains constant and acts like seeds of influence. In contrast, the opinion of listeners is unknown and will be estimated from their social interactions.

2 Methodology

2.1 Datasets

To build the datasets analyzed on this work,all the tweets were retrieved using the Twitter Streaming API. This API allows downloading tweets matching a set of keywords associated with each of the topics.

The final datasets are composed of 203,612 messages posted by 68,048 different users between December 11th and December 17th, 2017 in the case of Chileans elections and 36,090,661 messages written by 2,511,319 users from 15/09/2017 to 04/11/2017 in the Catalan independence issue.

2.2 Model to stimate user opinions

To infer the opinion of the users participating on the Twitter conversations we use the methodology introduced in [2]. We use a model based on the De Groot process that estimates opinions of users who interact on a social network from a minority of hubs whose opinion is known. In the model we have two types of users, elite and listeners.

The model assumes that we know with certainty the opinion of the elite.Thus, elite users have a fixed opinion that remains constant and act like seeds of influence. Elite users must have a strong constant opinion because we will keep it fixed. A user that frequently participates in the conversation can be considered to be engaged in the subject and, consequently, to have a well defined opinion. On the other hand, as the elite users will be seeds of influence, we need them to be relevant in the network, that is, they have a large number of retweets. We calculate the community structure of the most active users with a large number of retweets received and we analyzed the profile of each user. Finally we have assigned 64 elite nodes with the value 1 and 54 with -1 in the case of the Chilean elections and 184 with 1 and 139 with -1 in the independence of Catalonia issue.

The rest of the nodes are listeners. Initially their opinion is neutral and will be iteratively updated as the mean opinion of her incoming neighbors:

$$X_i(t+1) = \frac{\sum_j A_{ij} X\,j(t)}{k_i^{out}}, \tag{1}$$

where A_{ij} represents the elements of the retweet network adjacency matrix, which is 1 if and only if there is a link from j to i, and k_i^{out} corresponds to her out degree. The process is repeated until all nodes converge to their respective X_i value, lying in the range $-1 \leq X_i \leq 1$. Thus, the results of the model are given in a density distribution of nodes' opinion values $P(X)$.

2.3 Opinion Polarization

We will measure the political polarization of the conversation from the resulting density distribution of nodes opinion values $P(X)$. The polarization is given by [2]:

$$\rho = (1 - |\Delta A|)\, d \tag{2}$$

where

$$|\Delta A| = |P(X > 0) - P(X < 0)| \tag{3}$$

and $2d$ is the distance between positive and negative average opinions.

If $\rho = 1$ the distribution is perfectly polarized. In this case the opinion distribution function is two Dirac delta centered at -1 and $+1$ respectively. Conversely, $\rho = 0$ means that the opinions are not polarized at all.

3 Results

In the case of Chileans elections, we focus on analyzing the political polarization that emerges and tracking its evolutions during the week preceding the elections and the

final voting day. To this end, we first estimate the opinion of Twitter users from a minority of elite users, whose opinion was known. Next, we measure the resulting political polarization and analyze its evolution during that week (see fig.1). We find a shift on the opinions of users and the political polarization on the voting day. We explore to which extent this change on the behavior is explained by the engagement of new users commenting on the elections just that day or because users changed their minds during the last day. Finally, we show that the increase of the polarization observed on the previous day to the election is explained by a propaganda behavior of users who were already engaged to the conversation. However, the decrease in the polarization observed on the election day was caused by new users not so engaged to the political debate that entered the conversation acting as bridges between the two sides.

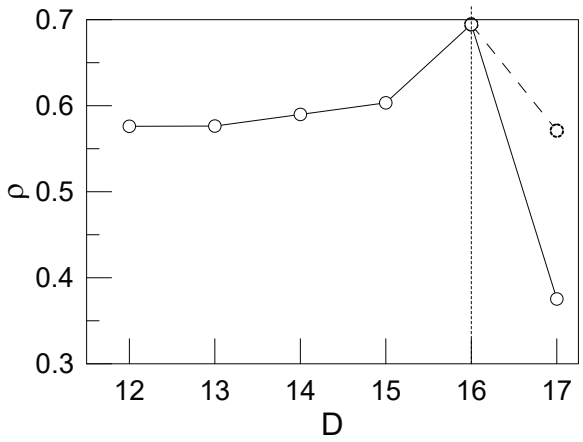


Fig. 1. Evolution of the probability distribution of the polarization index,ρ, for the retweet networks starting on December 11th until day D of 2017 Chilean Presidential elections. In dotted line, the values corresponding to the electoral campaign period 11-17December 2017 without taking into account the users who only participated on the voting day.

In the Catalan independence issue, we analyze polarization through tweets about the topic published in the period between 09/15/2017 and 03/11/2017. During this period important events occurred, the most relevant being the celebration of a referendum on independence not approved by the Spanish Government. The resulting distributions present a bimodal character with a small intermediate third pole, what shows a less polarized society, with individuals with not so extreme opinions. We find that the more active, engaged and influential users hold more extreme positions.

References

1. Olivares G., Crdenas J. P., Losada J. C., J. Borondo: Opinion Polarization during a Dichotomous Electoral Process. Complexity 2019, 5854037 (2019)
2. Morales A. J., Borondo J., Losada J. C., Benito R. M.: Measuring political polarization: Twitter shows the two sides of Venezuela. Chaos 25, 033114 (2015)

Bias in Social Interactions and Emergence of Extremism in Complex Social Networks

Vu X. Nguyen[1] and Gaoxi Xiao[1,2]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
[2] Complexity Institute, Nanyang Technological University - 18 Nanyang Drive, Singapore 637723, Singapore,
egxxiao@ntu.edu.sg

1 Introduction

Opinion dynamics is among the most important topics of complex network sciences. Extensive studies have been conducted to build mathematical models describing human interactions based on certain existing theories in sociology, social psychology, and complex sciences, e.g., balance theory [1], social comparison theory [2], cognitive dissonance theory [3], and social power theory [4], etc.

Emergence of extremism has been observed as a part of opinion dynamics in many social systems. To provide an explanation to such observations, the notion of *extreme agents* who intrinsically maintain a low susceptibility to persuasion and high persistence of opinion is typically presented. Such agents are also labeled as zealots, extremists, or inflexibles interchangeably. A few existing studies demonstrate that the presence of extremists persisting on long-lasting attitudes may swap a large part or even a whole of the population, causing a social network, with an open conflict, to end up being separated into different or even opposing opinion communities [5].

Motivated by observations that (i) assuming the pre-existence of extreme opinions does not explain how they emerge at the first place; and (ii) the existence of hatred may not necessarily be the force driving the emergence of extremism, we propose a new approach that allows extremism to emerge from largely "normal" actions that are less drastic than hatred or violence, which may appear to be quite similar to consensus making in the classic bounded-tolerance models, with the only difference that the pairwise consensus making could be slightly or significantly biased towards the two ends of the opinion spectrum. That is, instead of agreeing on a central value between two opinions in consensus making as that in the classical Deffuant model [6], we let right-wing opinion holders agree on an opinion that is to a certain extent biased towards right, while left-wing opinion holders' consensus tends to be somewhat biased towards left. It is shown that under such case, the normal consensus making, with a certain level of *bias* which may arguably be a part of human nature, may allow the emergence or even prevalence of extremism. This may help explain why extremism ideas can be observed in almost any human societies.

Our studies then further consider the effects of a few factors that arguably may be observed in many social systems. Due to length limit, only the results for one of them shall be reported in this abstract. That is, extremists tend to be less tolerant of different

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

opinions [7, 8]. We show that the factor, in fact, offers moderate opinions a better chance to survive and contributes to significantly dwarf the size of the extremist communities rather than helping them to prevail. Such observations, to a certain extent, may help explain why, though extremism widely exists in most social systems, extreme opinions seldom prevail to become the mainstream opinions of human societies.

2 Results

We consider a social network whose opinions varies from 0 (extremely hate) to 1 (extremely like). In a network with a certain tolerance range identical across the entire population, it is shown that in the absence of bias, all agents converge to a few major opinion clusters over time. Under the effects of bias, however, the major clusters tend to shift towards the two ends of the opinion axis. The bias in local interactions leads to the emergence and persistent existence of the clusters holding extreme opinions, and concurrently the decline of clusters holding moderate opinions (see Fig. 1).

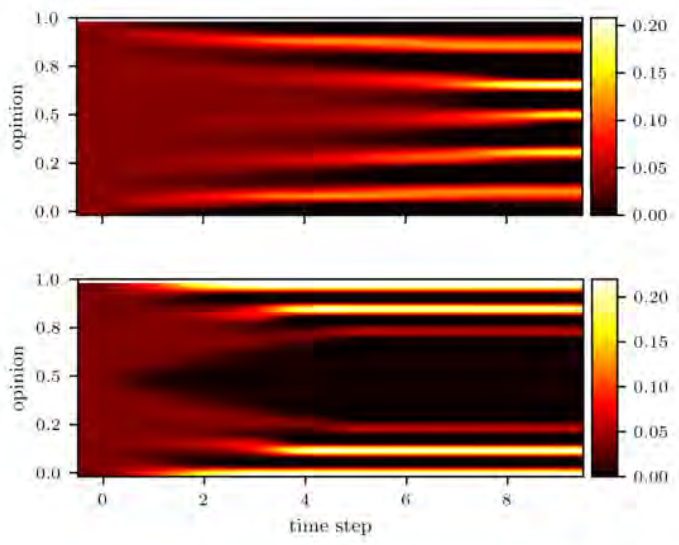


Fig. 1. Illustration of temporal distributions of opinions of 20000 agents connected into an ER network with an average nodal degree of 20 without (top) and with (bottom) local bias. Each time step at which the temporary opinion distribution is recorded corresponds to 20000 pair-wise interactions.

We then formulate the tolerance range as a function that reaches its maxima and minimas at the neutral opinion (opinion 0.5) and the two extreme opinions (opinion 0 and 1), respectively. Figure 2 shows that regardless of the bias factor, the peaks closer to the two ends of the opinion axis quickly form up while moderate opinions keep evolving. This is largely due to the narrower tolerance ranges of extreme agents, making their

communities stop adopting newcomers and maintain their current sizes. The extremists' low tolerances offer a higher chance for their persistence but diminish their chance to grow in popularity. The observations may provide an explanation to why communities of extremists can emerge and persist widely in many real-life societies yet rarely prevail.

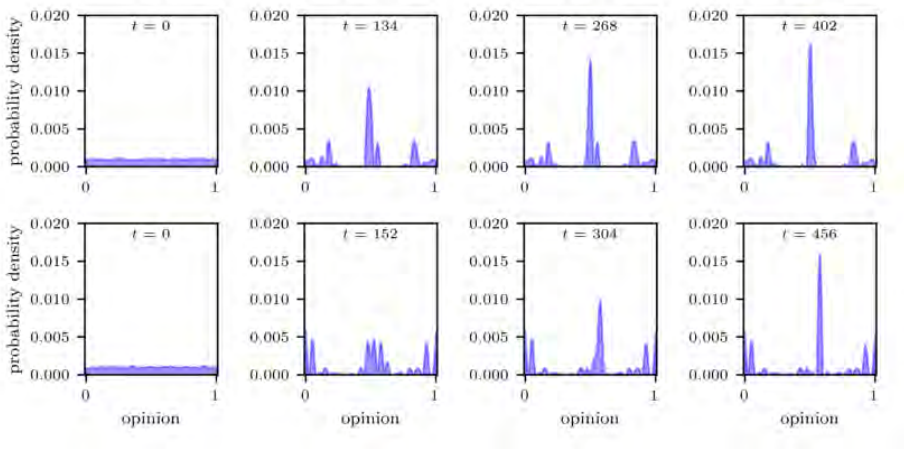


Fig. 2. Time evolution of opinions in a Facebook ego network of 4039 nodes without (top row panels) and with (bottom row panels) bias. The tolerance range is largest at the centrist opinion and linearly narrowed down as the opinion shifts towards the two extremities. Each time step at which the opinion distribution is recorded corresponds to 4039 pair-wise interactions.

Acknolwledgement: This work is partially supported by Ministry of Education, Singapore, under contract MOE2016-T2-1-119.

References

1. Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological Review*, 63(5):277, 1956.
2. Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954.
3. Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford University Press, 1957.
4. John RP French Jr. A formal theory of social power. *Psychological Review*, 63(3):181, 1956.
5. Gérard Weisbuch, Guillaume Deffuant, and Frédéric Amblard. Persuasion dynamics. *Physica A: Statistical Mechanics and its Applications*, 353:555–575, 2005.
6. Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.
7. James O Whittaker. Perception and judgment in the political extremist 1. *Journal of Communication*, 17(2):136–141, 1967.
8. Jan-Willem van Prooijen and André PM Krouwel. Extreme political beliefs predict dogmatic intolerance. *Social Psychological and Personality Science*, 8(3):292–300, 2017.

Role of Facebook in Building a Learning Community: Case of Japanese Study Abroad Program

Hiromitsu Goto[1], Yuji Nakatani[2], and Chikara Funabashi[3]

[1] Turnstone Research Institute, Inc., Japan,
hiromitsu.goto.phys@gmail.com,
[2] Real College Inc., Japan
[3] Asian Leaders Connecting Hub Pte. Ltd., Singapore

1 Introduction

Online social networks provide platforms for communication, interaction and collaboration between friends. They allow us to expect the realization of the effective learning support using online communities. However, it is necessary to further understand the relationship between both online and offline social behavior as discussed in [1].

The Tobitate! Young Ambassador Program is a Japanese public-private partnership aimed at helping Japanese students study abroad [2], launched in 2014. The program recruits students twice a year to support studying abroad, and the students adopted in the program will study abroad after the joint pre-training as TOBITATE students, moreover, the joint post-training is conducted after the study abroad. The desired result is for the students to form a learning community through a common program, and for the program to build a globalized human resources development community.

In this work, we evaluate the learning support in the TOBITATE program through analysis of Facebook friendship network combined with student data from the program. We focus on the Facebook Group [3], which is reported as main online community for the participants by the alumni association of the program. The Facebook Group participants include college students, office members who manage the program and small number of high school students who we ignore here. In order to examine network characteristics, we use the following participants' attributes;

- **Class** that represents when the participant joined the program. We use numbers corresponding to the student's participation period as labels. The office member are labeled as zero.
- **Course** selected by the student in the program. They depend on the plan of study abroad such as A) natural science, B) unique challenge, C) world-leading institution, D) emerging economies and E) regional development.
- **Prefecture** where the university to which the student belongs originally is located. There are 47 prefectures in Japan, and we assign the numbers in order from the north. For example, 13th prefecture represents Tokyo.
- **Country** where the student mainly studied abroad.

We use only public available friendship data on the individual Facebook pages and the personal information for the nodes of TOBITATE students are provided by the Japanese Student Services Organization (JASSO).

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

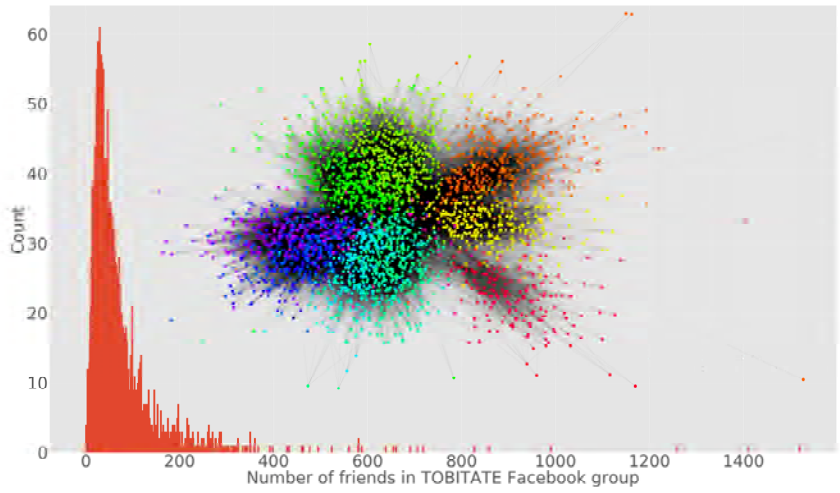


Fig. 1. The distribution of friendship links in TOBITATE Facebook group and visualization of friendship network. The color of dots represent classes of participants. The purple, blue and red colors correspond to staff members, 1st and 10th TOBITATE students, respectively.

2 Results

We found that Facebook play an important role in forming a small-world network as the learning community for participants of TOBITATE program. The network consists of 3,265 participants and 111,977 friendship ties as of April 11, 2019. There are 45 office member to manage the program and 3,220 TOBITATE students up to 10th generation. The network has properties of small-world network, whose average shortest path length for 2.33 and the average clustering coefficients for 0.247. We show its degree distribution and the visualization in Fig. 1, where the color of dots represent classes of participants. It is clear from the visualization result that there are communities for classes. As shown in Fig. 2, the friendship ties between different attribution types increase as the size of these end nodes except for the class case. In the class case, moreover, the students of similar generations tend to have friendship relations and the office members are connected equally to all classes. This is because the program provides opportunities for similar classes to form a learning community through the joint pre- and post-training to study abroad in offline. As expected, therefore, this results show that offline contact opportunities are helping to form online friend relationships, and program office members contribute to build the learning community with properties of small-world network. This fact also appears that the office members have higher betweenness centrality than the other classes' one.

Summary. There remains a need for practical applications of online social network analysis, such as student learning support. We have evaluated the learning support in the Japanese study abroad program, which is named as TOBITATE program, through analysis of Facebook friendship network combined with student data from the program.

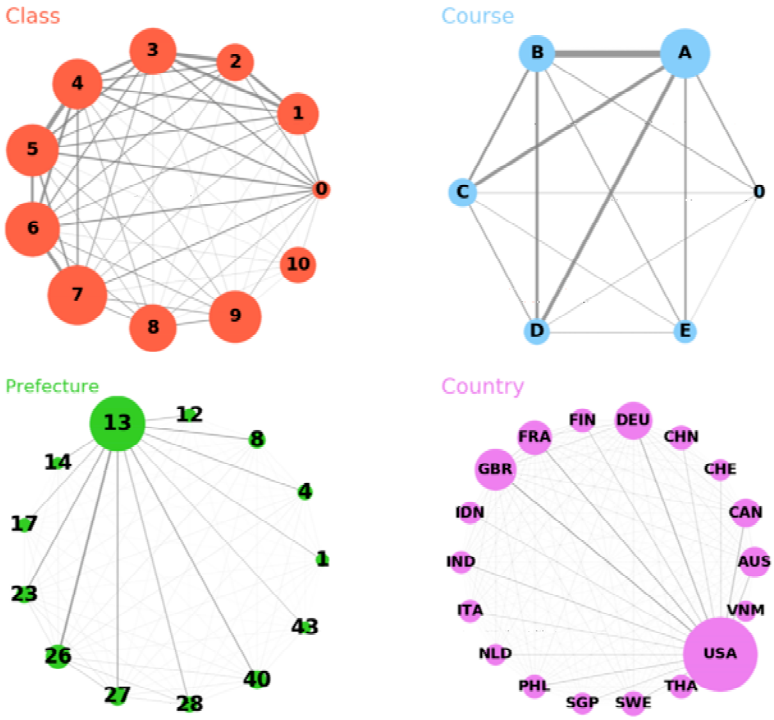


Fig. 2. Networks of node's attribution. The attribute types are depicted as nodes, and their size is scaled to the size of their corresponding attribute types. We only show the nodes having more fifty size in below two networks. The width is proportional to the total number of their links and self-loops are removed.

We have observed that the offline contact opportunities formed online friendships and that the support of communicators helped to build a learning community across the generations. Our study contributes to realize the effective learning support using the online community.

Acknowledgments

This work is supported in part by investigating projects that contribute to the promotion of student support (JASSO Research).

References

1. Hristova, D. et al. "Keep your friends close and your facebook friends closer: A multiplex network approach to the analysis of offline and online social ties." Eighth International AAAI Conference on Weblogs and Social Media. 2014.
2. Tobitate! Study Abroad Initiative, URL https://www.tobitate.mext.go.jp/about/english.html
3. TOBITATE Facebook Group, URL https://www.facebook.com/groups/1585059971706689/

*Everything You Always Wanted to Know About AI**

(*But Were Afraid to Ask)

Nowcasting Digital Skills with Wikipedia

(Working Paper - Version 16/9/19)

Fabian Stephany[1 2 3]

[1] Humboldt Institute for Internet and Society, Französische Straße 9, 10117 Berlin, Germany
[2] Wittgenstein Centre, Welthandelsplatz 2, 1020 Vienna, Austria
[3] AtomLeap GmbH, Oranienstraße 183, 10999 Berlin, Germany
OrcidID: 0000-0002-0713-6010, mail@fabianstephany.com

1 Introduction

Digital technologies have a pervasive effect on our society. They augment or transform various previously analogue processes of value creation, capture, and exchange [1,2]. Hence, on the labour market, the skillful development and application of relevant digital technologies are in strong demand. However, early research findings indicate that the labour demand of certain tech industries is not met. The talent pool does not grow at the pace of industry demand and precise skill requirements related to growing technologies, such as Artificial Intelligence (AI), remain opaque [3].

This work proposes a network perspective in order to empirically identify the relevant ICT skills related to AI, how their composition changes over time, and how they could be predicted with online data. With the example of the US tech sector, two data sources are employed: The US' most popular online tech-job platform *dice.com* allows to relate ICT skills, in a network structure, from an industry perspective. Two skills are connected in an industry demand network, if they are jointly required by the same job advertisement. In addition, data from the online encyclopedia *Wikipedia* allows to create a network on the online knowledge side of ICT skills[4]. Here, skills are related in a network, if their respective articles are connected by a reference hyperlink (Figure 2 in the Appendix illustrates.

Similarly, past skill networks can be constructed: Information about previous job advertisements is stored in older versions of *dice.com* on the web-archive[5] since 2004 and *Wikipedia's* reference history allows to reconstruct article networks at any point in time during the history of the encyclopedia. Over time, composition developments of the skill networks can be compared between and within the two network environments. The underlying hypothesis is that developments in the online knowledge network (*Wikipedia*) proceed changes in the industry demand network (*dice.com*). While it is one of the main challenges of this work to assess the validity of this hypothesis, previous studies have shown that the edit activities of the *Wikipedia* crowd enable early

[4] It is reasonable to assumed that *Wikipedia* article editors possess relevant topical knowledge on the digital skill the edited article is about.

[5] http://web.archive.org/web/*/http://www.dice.com/jobs

COMPLEX NETWORKS 2019

The 8th International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

predictions on movie sales [4], electoral popularity [5], stock prices [6], knowledge hubs [7] or even global spreading of diseases [8].

2 Results

For the first exploratory part of the research, two ego-centred networks are regarded: All job postings advertised with the tag "Artificial Intelligence"[6] on *dice.com* on September 14th, 2019, are considered. Skill tags are connected as nodes in a network, if they appear in the same advertisement. Similarly, all articles linked with the *Wikipedia* article "Artificial Intelligence"[7] and the articles they refer to are connected in a network[8]. For both networks, a common set of most relevant overlapping skill tags are identified[9]. With the use of the *Wikipedia* edit history and the web-archive, the date, when these tags entered the respective ego-centred AI-networks on both *Wikipedia* and *dice.com* are registered[10]. On average, skill nodes joined the *Wikipedia* AI-network 15 weeks before they appeared in relation to AI on *dice.com*, as illustrated in Figure 1.

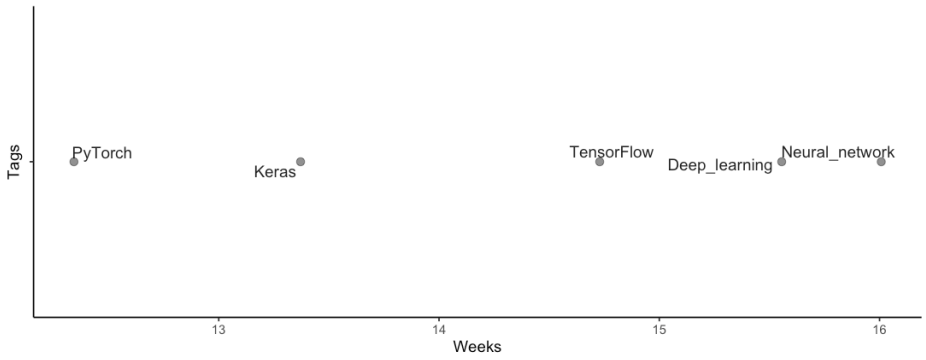


Fig. 1. *In both ego-networks of AI (Wikipedia and dice.com), new nodes join over time. However, nodes entered significantly earlier on Wikipedia. For a set of selected skills, on average, their articles have been linked to "Artificial Intelligence" about 15 weeks before they had been announced on job advertisements about AI.*

With this first indication of the predictive potential of *Wikipedia* data on digital skills, future extensions of this work focus on an identification of AI-cliques by clus-

[6] https://www.dice.com/jobs?q=Artificial+intelligence&l=

[7] https://en.wikipedia.org/wiki/Artificial_intelligence

[8] Nodes can, at most, have a distance of one iteration from the original AI article.

[9] ApacheMX NeT, Big data, Caffe, Computer vision, Data mining, Data science, Deep learning, Keras, Machine learning, Natural language processing, Neural networks, Predictive analytics, Python, PyTorch, TensorFlow, Theano, R, RNN

[10] For five of the key skills it was possible to find the entry into the respective networks.

tering algorithms and under the consideration of all relevant skill tags that are currently in use on *dice.com* and their respective *Wikipedia* articles. Moreover, future extension take centrality metrics, e.g., *Eigenvector centrality*, into account for comparing the state and development of skill nodes in both network environments. Lastly, the comparison of network similarities, e.g., *Jaccard similarity*, allows future investigations to evaluate the development of the AI-cliques within and across the two network environments. In addition, following investigations should take regional digital knowledge geographies [9] into account, too.

Overall, the insights of this project can support businesses in developing a data-driven strategy for the acquisition and the development of adequate skills needed to implement and leverage new technologies at best. Furthermore, the empirical relationship of digital skill sets will help to establish a common taxonomy to be used by policy makers, education providers, and recruiters, so that job market mismatches can be reduced. Lastly, a potential predictive power of the online knowledge network could help to develop farsighted programmes for the training of digital skills in the future.

Summary. With the use of online data from the tech job platform *dice.com* and the online encyclopedia *Wikipedia*, two networks of digital skills are created around the topic of Artificial Intelligence. Initial research indicates that new skill tags first join the *Wikipedia* network, before they appear in AI-related job announcements on *dice.com*. The findings of this work could be used in order to create a data-driven strategy for the acquisition and the development of adequate skills needed to implement and leverage new technologies at best.

References

1. Yoo, Youngjin, Ola Henfridsson, and Kalle Lyytinen. "Research commentarythe new organizing logic of digital innovation: an agenda for information systems research." Information systems research 21, no. 4 (2010): 724-735.
2. Nambisan, Satish, Kalle Lyytinen, Ann Majchrzak, and Michael Song. "Digital Innovation Management: Reinventing innovation management research in a digital world." Mis Quarterly 41, no. 1 (2017).
3. De Mauro, Andrea, Marco Greco, Michele Grimaldi, and Paavo Ritala. "Human resources for Big Data professions: A systematic classification of job roles and required skill sets." Information Processing & Management 54, no. 5 (2018): 807-817.
4. Mestyn, Mrton, Taha Yasseri, and Jnos Kertsz. "Early prediction of movie box office success based on Wikipedia activity big data." PloS one 8, no. 8 (2013): e71226.
5. Yasseri, Taha, and Jonathan Bright. "Can electoral popularity be predicted using socially generated big data?." it-Information Technology 56, no. 5 (2014): 246-253.
6. Moat, Helen Susannah, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. "Quantifying Wikipedia usage patterns before stock market moves." Scientific reports 3 (2013): 1801.
7. Stephany, Fabian, and Fabian Braesemann. "An exploration of wikipedia data as a measure of regional knowledge distribution." In International Conference on Social Informatics, pp. 31-40. Springer, Cham, 2017.
8. Generous, Nicholas, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. "Global disease monitoring and forecasting with Wikipedia." PLoS computational biology 10, no. 11 (2014): e1003892.

9. Stephany, Fabian, Fabian Braesemann, and Mark Graham. ”Coding Together-Coding Alone: The Role of Trust in Collaborative Programming.” (2019).

3 Appendix

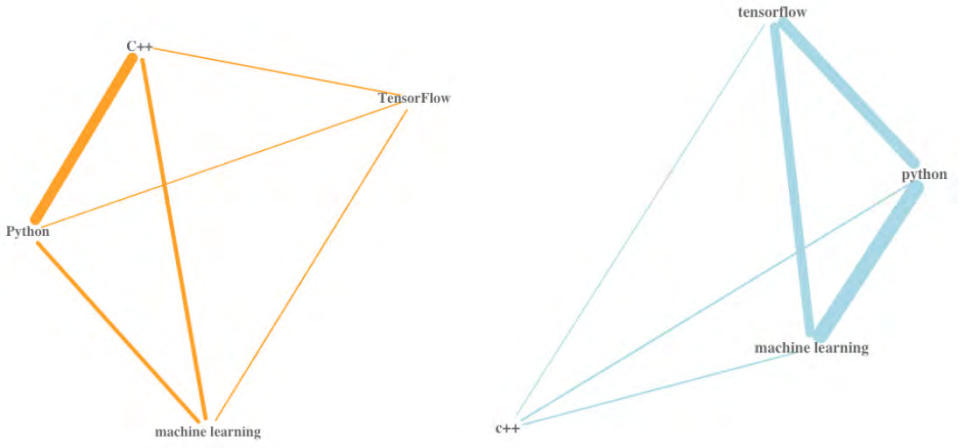


Fig. 2. *Two skill networks are constructed with the example of Wikipedia articles (lhs) and dice.com job advertisements (rhs). The stronger the edges between the nodes, the more references articles share with each other / the more jobs have jointly advertised the same skills. The example of the nodes, most strongly connected to the software TensorFlow, serves as an illustration. TensorFlow belongs to the field of machine learning skills and was written in the programming languages Python and C++. Considering the subgraph of these four tags, exclusively, one can see that C++ is more strongly connected in the online knowledge network than in the industry demand network. This could potentially be explained by the fact that C++ is less relevant for the application of TensorFlow, than it was for the development of the software.*

Part XIX

Synchronization, Resilience and Control

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Complete Networks: Discontinuous Dynamics, Information Invariants and Synchronization

J. Leonel Rocha[1] and S. Carvalho[2]

[1] CEAUL. ADM, ISEL-Eng. Superior Institute of Lisbon, IPL
Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
jrocha@adm.isel.pt
[2] CEAFEL. ADM, ISEL-Eng. Superior Institute of Lisbon, IPL
Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
scarvalho@adm.isel.pt

1 Introduction and preliminars

Several authors have dedicated their investigation to the study of the information theory and its applications. The amount of information produced by a network may be measured by the mutual information rate. This measure together with the Kolmogorov-Sinai entropy are expressed in terms of the conditional Lyapunov exponents. Also, it is well known that chaotic systems can be synchronized. The recognized potential for communications systems has driven this phenomenon to become a distinct subfield of nonlinear dynamics, see [1], [2], [3] and [6]. Information theory and synchronization are directly related in a network. Motivated by the theoretical and practical connection between the information invariants (or measures) and the phenomenon of synchronization, our purpose in this work is to analyze the relations between the mutual information rate, the Kolmogorov-Sinai entropy and the synchronization in the space of complete networks of order N. The networks topologies are characterized by circulant matrices and its conditional Lyapunov exponents are explicitly determined. For different types of discontinuous local dynamics, necessary and sufficient conditions for the occurrence of synchronization with or without the negativity of the conditional Lyapunov exponents are presented. Some properties of the mutual information rate and the Kolmogorov-Sinai entropy are established, depending on the topological entropy of the individual chaotic nodes and on the synchronization interval. The novelty of these results is established in comparison with the studies presented in [2], [3] and [6].

Consider a network of N identical chaotic dynamical oscillators or units, described by a connected and unoriented graph $G=(V,E)$, with no loops and no multiple edges. In each node the dynamic of the oscillators is defined by $\dot{x}_i = f(x_i)$, with $f:\mathbb{R}^n \to \mathbb{R}^n$ and $x_i \in \mathbb{R}^n$ is the state variables of the node i. Throughout this work we will consider the space of complete network of order N with $\frac{N(N-1)}{2}$ edges, will be denoted by K_N. Notice that every vertex of G has degree $N-1$. Consider A the adjacency matrix of K_N and $D = diag(N-1,\ldots,N-1)$, then $L=[l_{ij}]=A-D$ represents the laplacian matrix of the complete graph and is written in the following form,

$$L=\begin{bmatrix} -(N-1) & 1 & 1 & \ldots & 1 \\ 1 & -(N-1) & 1 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 1 & -(N-1) \end{bmatrix}.$$

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

The matrix L has exactly two eigenvalues $\mu_1 = 0$, a simple root, and $\mu_2 = -N$, with multiplicity $N-1$. The dynamics of these N coupled oscillators can be expressed by the following system of differential equations:

$$\dot{x}_i = f(x_i) + \sigma \sum_{j=1}^{N} l_{ij} x_j, \tag{1}$$

where $i = 1, 2, ..., N$, $\sigma > 0$ is the coupling parameter, see, [1], [3], [4] and [5]. Let f' be the derivative of f, then the jacobian matrix of this network K_N has also two eigenvalues $\lambda_1 = f'$, also a simple root, and $\lambda_2 = f' - N\sigma$, with multiplicity $N-1$ and is written as follows,

$$J = \begin{bmatrix} f'-(N-1)\sigma & \sigma & \dots & \sigma \\ \sigma & f'-(N-1)\sigma & \dots & \sigma \\ \dots & \dots & \dots & \dots \\ \sigma & \sigma & \dots & f'-(N-1)\sigma \end{bmatrix}.$$

Every matrix associated with a complete network K_N has a certain regularity, both matrices L and J are circulant matrices, so they are diagonalizable and have the same eigenspaces. In this context the following results are proved and numerical studies are included in [5].

2 Local dynamics: discontinuous piecewise linear maps $s > 1$

In this section we consider the space of all the complete networks K_N, given by Eq.(1), where the local dynamics in each node is defined by $f : I \subset \mathbb{R} \to \mathbb{R}$, a discontinuous piecewise linear map with constant slope $s > 1$ everywhere. We consider the following parameters space $\Sigma^+ = \left\{(N, s, \sigma) \in \mathbb{R}^3 : N \in \mathbb{N} \setminus \{1\}, s > 1, \sigma > 0\right\}$.

Property 1. Consider the (K_N, Σ^+) space. Let $f : I \to \mathbb{R}$ be a discontinuous piecewise linear map with slope $s > 1$ everywhere. The synchronization interval of K_N is given by

$$\sigma_1 = \frac{s-1}{Ns} < \sigma < \frac{s+1}{Ns} = \sigma_2. \tag{2}$$

The chaoticity of the local dynamics is measured by the topological entropy of f, i.e., $h_{top}(f) = \log|s|$.

Proposition 1 *Consider the (K_N, Σ^+) space. Let $f : I \to \mathbb{R}$ be a discontinuous piecewise linear map with slope $s > 1$ everywhere, I_σ be the synchronization interval, given by Eq.(2), and $I_{\lambda_\perp^-}$ be the interval where $\lambda_\perp \leq 0$. It is verified that:*

(i) $I_\sigma \cap I_{\lambda_\perp^-} \neq \emptyset$ *if and only if* $1 < s < 1 + \sqrt{2}$;
(ii) $I_\sigma \cap I_{\lambda_\perp^-} = \emptyset$ *if and only if* $s \geq 1 + \sqrt{2}$.

Proposition 2 *Consider the (K_N, Σ^+) space. Let $f : I \to \mathbb{R}$ be a discontinuous piecewise linear map with slope $s > 1$ everywhere, I_σ be the synchronization interval, given by Eq.(2), and $I_{\lambda_\perp^-}$ be the interval where $\lambda_\perp < 0$. It is verified that:*

(i) for $1 < s < 1 + \sqrt{2}$,
  (a) if $\sigma \in I_\sigma^- = I_\sigma \cap I_{\lambda_\perp^-}$, *then* $I_C = H_{KS}$;
  (b) if $\sigma \in I_\sigma^+ = I_\sigma \setminus I_\sigma^-$, *then* I_C *increases and* H_{KS} *decreases;*
(ii) if $s \geq 1 + \sqrt{2}$, *then* I_C *increases and* H_{KS} *decreases, with* $I_C \neq H_{KS}$, $\forall \sigma \in I_\sigma$.

3 Local dynamics: discontinuous piecewise linear maps $|s| > 1$

Through this section we study complete networks K_N, where the local chaotic dynamics are defined by $f : I = [b_1, b_2] \subset \mathbb{R} \to \mathbb{R}$, a discontinuous piecewise linear map, with constant slope $|s| > 1$ everywhere, where $\Sigma^{\pm} = \left\{(N, s, \sigma) \in \mathbb{R}^3 : N \in \mathbb{N} \setminus \{1\}, |s| > 1, \sigma > 0\right\}$. Let r_1 be the only positive real root of the polynomial $s^4 - 2s - 1 = 0$ and r_2 be the only positive real root of the polynomial $s^4 - 2s^2 - 2s - 1 = 0$. Notice that $1 < r_1 < r_2$.

3.1 Equal amplitudes of the subintervals with slope $s > 1$ (a^+) and slope $s < -1$ (a^-)

Proposition 3 *Consider the $(K_N, \Sigma^{\pm})$ space. Let $f : I \to \mathbb{R}$ be a discontinuous piecewise linear map with slope $|s| > 1$ everywhere, I_σ be the synchronization interval, given by Eq.(2), and $I_{\lambda_\perp^-}$ be the interval where $\lambda_\perp < 0$. For $a^+ = a^-$, it is verified that:*

(i) $I_{\lambda_\perp^-} \subset I_\sigma$ *if and only if* $1 < |s| < r_1$;

(ii) $I_\sigma \cap I_{\lambda_\perp^-} \neq \emptyset$ *if and only if* $r_1 \leq |s| \leq r_2$;

(iii) $I_\sigma \cap I_{\lambda_\perp^-} = \emptyset$ *if and only if* $|s| > r_2$.

Proposition 4 *Consider the $(K_N, \Sigma^{\pm})$ space. Let $f : I \to \mathbb{R}$ be a discontinuous piecewise linear map with slope $|s| > 1$ everywhere, I_σ be the synchronization interval, Eq.(2), and $I_{\lambda_\perp^-}$ be the interval where $\lambda_\perp < 0$. For $a^+ = a^-$ and $1 < |s| < r_1$, it is verified that:*

(i) *if* $\sigma \in I_{\lambda_\perp^-}$, *then* $I_C = H_{KS}$;

(ii) *if* $\sigma \in I_\sigma \setminus I_{\lambda_\perp^-}$ *and* $\sigma_1 < \frac{\sqrt{s^2-1}}{N}$, *then I_C increases and H_{KS} decreases, with* $I_C \neq H_{KS}$;

(iii) *if* $\sigma \in I_\sigma \setminus I_{\lambda_\perp^-}$ *and* $\sigma_2 > \frac{\sqrt{s^2+1}}{N}$, *then I_C decreases and H_{KS} increases, with* $I_C \neq H_{KS}$.

3.2 Different amplitudes of the subintervals with slope $s > 1$ and slope $s < -1$

Proposition 5 *Consider the $(K_N, \Sigma^{\pm})$ space. Let $f : I \to \mathbb{R}$ be a discontinuous piecewise linear map with slope $|s| > 1$ everywhere. Consider the measures I_C and H_{KS}, with $a^+ \neq a^-$. If $I_C = H_{KS}$, then $|s - N\sigma| < 1$ and $a^+ > a^-$.*

Acknowledgments: *Research funded by the project IPL - MISRedes, IDI&CA 2019, FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the projects UID/MAT/00006/2019 (CEAUL), UID/MAT/04721/2019 (CEAFEL) and ISEL.*

References

1. Boccaletti, S., Kurths, J., Osipov, G., Valladares, D.L., Zhou, C.S.: The synchronization of chaotic systems. Phys. Rep. 366, 1–101 (2002)
2. Cao, A., Lu, J.: Adaptive synchronization of neural networks with or without time-varying delays. Chaos 16, 013133 (2006)
3. Pecora, L.M., Carroll, T.L.: Driving systems with chaotic signals. Phys. Rev. A 44, 2374–2383 (1991)
4. Rocha, J.L., Caneco, A.: Mutual information rate and topological order in networks. Chaotic Modeling and Simulation, Int. J. Nonlinear Science 4, 553–562 (2013)
5. Rocha, J.L., Carvalho, S.: Information Measures and Synchronization in Complete Networks. submitted (2019)
6. Shuai, J.W, Wong, K.W., Cheng, L.M.: Synchronization of spatiotemporal chaos with positive conditional Lyapunov exponents. Phys. Rev. E 56, 2272 (1997)

Strength optimization of materials with complex microstructure: Beam Network Model

Seyyed Ahmad Hosseini, Paolo Moretti, and Michael Zaiser

Institute of Materials Simulation (WW8), Friedrich-Alexander-Universitt Erlangen-Nrnberg (FAU), Dr.-Mack-Str. 77, 90762 Frth, Germany,
ahmad.hosseini@fau.de,
WWW home page: http://www.matsim.techfak.uni-erlangen.de/

1 Introduction

Beam network models (BNM) are extensively used to model fracture of materials subject to external stresses, especially in cases in which the material at hand exhibits a complex microstructure. A typicial example of complex micorstructure is that encountered in fibrous materials such as bone, which consist of a complex and multi-scale network of fibers and cross-links. The edges of such a networks can be treated as beams, load carrying elements that deform and break [1]. BNM, as opposed to even more simplified models such as random fuse models, or random spring models, explicitly preserve fundamental features of continuum mechanics such as the tensorial nature of stress and strain, and the conservation of linear and angular momentum. This allow to 'tune' them to reproduce, in principle, macroscopic elastic properties of any type of material.

In order to ensure strength and reliability of materials that can be modelled as beam networks, design rules and paradigms need to be adapted to account for the statistical variability of materials properties. In the present investigation, we illustrate this for a simple design problem, namely the optimal configurations of a uni-axially strained bundle of load-carrying brittle fibers with variable amount of cross links. In the absence of structural disorder, this problem has a trivial solution which, however, turns out to be the worst possible solution if the system is large or the material is strongly disordered.

2 Results

When an axial load is applied to a fully connected network (FBN) (Figure 1), it is evident that initially the cross-link (CL) beams do not carry any load. The basic question we are asking is thus: Assuming that CL beams are associated with a deterioration in strength of the connecting load-carrying (LC) beams, what is the optimal degree of cross linking for a system of given size L (which defines the linear dimension of the network) and given Weibull exponent β of the strength distribution of the elementary beams (i.e., degree of material disorder).

We consider a situation where the introduction of cross links weakens the load carrying beams as the 'welds' connecting the beams weaken the beam structure (an example are chemically cross linked carbon nanotubes where the chemical cross links represent imperfections in the otherwise regular sp^2 bonding network, [2, 3].) In our model, we

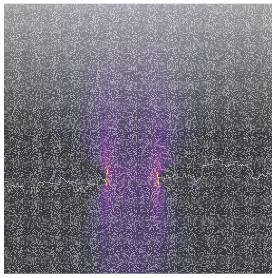


Fig. 1. A transversely propagating crack in BNM.

introduce such weakening by randomly choosing one of the four LC beams that connect to a newly introduced CL beam, and multiplying its strength by a factor $f_t \leq 1$. This introduces a trade-off between strength reduction of the LC beams, and a potential gain in strength due to cross linking. The resulting overall strength is, for different values of f_t and β, shown in Figure 2 versus the respective cross-link ratio.

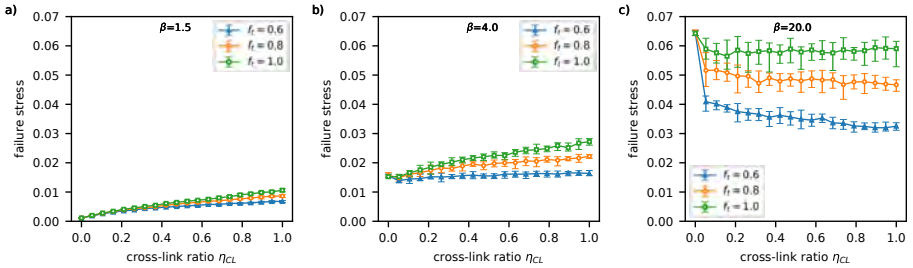


Fig. 2. Effect of cross-linking on mechanical strength of fiber bundles, when cross linking introduces damage into the load carrying beams. Strength vs. cross-link ratio, number of realizations $N = 20$ for each data point. After introducing each cross link, 1 of the 4 LC beams connecting to that link is weakened by reducing its strength by a factor f_t. Original beam strengths are Weibull distributed with shape parameters $\beta = 1.5, \beta = 4.0$ and $\beta = 20.0$, system size is $L = 256$.

In the limit of low disorder (Figure 2, right), it is evident from the figure that the introduction of cross links is unfavorable as even low cross link ratios reduce the overall strength by a factor close to f_t. Remarkably, a small knock-down effect on strength is manifest even for $f_t = 1$, indicating that a cross-linked bundle of fibers is here *weaker* than its unconnected counterpart. This can be understood from the different failure modes of unconnected fiber bundles and of connected beam networks: For a connected network, lateral load re-distribution leads to localized damage clusters which form as weaker-than-average beams fail in a correlated manner and that extend in *lateral* direction. In studies of random fuse networks, such damage clusters were shown to control system strength in a manner very similar to small cracks that become critical at the failure stress [4, 5]. Thus, we observe that in a material with low disorder, cross linking may reduce strength even if it does not introduce damage, because cross links here fa-

cilitate the formation and lateral propagation of a critical crack which is impossible as long as the system consists of unconnected fibers. The effect is, of course, exacerbated if in addition the cross links are associated with damage (reduction in strength) of the load carrying fibers.

In the high disorder limit, on the other hand, introduction of CL beams leads to a significant strength enhancement even if the associated damage to the LC beams is significant (Figure 2, left). In this limit, FBN made of intact beams ($f_t = 1$) are strongest with a strength that exceeds the strength of a bundle of isolated fibers by more than one order of magnitude. Structures where cross linking introduces damage ($f_t < 1$) are weaker by a factor close to f_t, but still much stronger than the corresponding fiber bundles. For intermediate disorder ($\beta = 4$), finally, we find an intermediate behavior where the benefit of cross linking depends on the amount of damage introduced (Figure 2, center): With $f_t = 1$ and $f_t = 0.8$, the FBN is strongest, whereas for $f_t = 0.6$ strength is almost independent on cross linking degree and for $f_t < 0.6$ the strongest structure is represented by an unconnected fiber bundle.

Summary. We use beam network models to find optimal configurations for fibrous, bio-inspired materials undergoing fracture. Our investigation shows that structural disorder of materials can have serious consequences for design considerations. Using a trivial example, namely a bundle of fibers carrying an axial load, we demonstrated that the optimal design for a perfect material may perform worst when built of components (beams) made from a strongly disordered and thus unreliable material. Disorder necessitates structural redundancy in form of the creation of alternative load transmission paths through cross linking, even if such cross links appear superfluous from elasticity calculations but incur a cost in terms of added weight or even in terms of reduced strength of the load carrying fibers.

Acknowledgments The authors acknowledge support by DFG under Grant No 1Za 171/9-1 and under 377472739/GRK 2423/1-2019 FRASCAL. Support by the European commission under H2020-MSCA-RISE project no. 734485 FRAMED is also gratefully acknowledged.

References

1. Manzato, C., Shekhawat, A., Nukala, P. K. V. V., Alava, M. J., Sethna, J. P., Zapperi, S.: Fracture strength of disordered media: Universality, interactions, and tail asymptotics. Phys. Rev. Lett. 108(6), 065504 (2012)
2. Moghadam, R. M., Hosseini, S. A., Salehi, M.: The influence of StoneThrowerWales defect on vibrational characteristics of single-walled carbon nanotubes incorporating Timoshenko beam element. Physica E 62, 80–89 (2014)
3. Yang, M., Koutsos, V., Zaiser, M.: Size effect in the tensile fracture of single-walled carbon nanotubes with defects. Nanotechnology 18(15), 155708 (2007)
4. Lennartz-Sassinek, S., Zaiser, M., Main, I. G., Manzato, C., Zapperi, S.: Emergent patterns of localized damage as a precursor to catastrophic failure in a random fuse network. Phys. Rev. E 87(4), 042811 (2013)
5. Zaiser, M., Lennartz-Sassinek, S., Moretti, P.: Crack phantoms: localized damage correlations and failure in network models of disordered materials. J. Stat. Mech. 8, P08029 (2015)

Predicting collapse of adaptive networked systems without knowing the network

Leonhard Horstmeyer[1,2], Tuan Minh Pham[1,2], Jan Korbel[1,2], and Stefan Thurner[1,2,3,4]

[1] Section for the Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Vienna, Austria
[2] Complexity Science Hub Vienna, Josefstädterstrasse 39, A-1080 Vienna, Austria
[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[4] IIASA, Schlossplatz 1, 2361 Laxenburg, Austria

1 Introduction

Networked dynamical systems are ubiquitous in nature and man-made systems. They appear in food-webs, eco-systems, financial markets, communication networks, epidemiology, etc. Often the stability and resilience of such systems depend on the details of the underlying interaction networks. To predict collapse of networked dynamical systems is a challenging task, because often networks are large and cannot be observed directly. Therefore, early warning signals for such systems practically do not exist. We show that as a consequence of a mathematical theorem it is possible to detect the last stages before a crash by observing a quantization effect in the components of the system, without knowing the network structure. This allows us to predict the collapse of various networked dynamical systems.

2 Results

Let us consider a minimal version of a networked dynamical system described by

$$\frac{d}{dt}X_i(t) = \sum_{j=1}^{N} M_{ij}X_j(t) - \Phi X_i(t) \quad , \tag{1}$$

where M_{ij} is the interaction matrix and Φ is decay rate. Let us consider a binary matrix without self-interactions, i.e., $M_{ij} = \{0,1\}$ and $M_{ii} = 0$. One can show that there are two distinct phases. First, when the interaction matrix contains no cycle, then for $\Phi > 0$ are all X_i decay to zero for $t \to \infty$. On the contrary, if M_{ij} contains cycles, then some X_i remain positive. This fact is the basic mechanism of all autocatalytic systems. We present a theorem, a corollary of the famous Perron-Frobenius theorem which can distinguish between network with the one remaining cycle and network with more cycles without knowing the network topology.

Theorem 1 (Eigenvector Quantization) *.Let G be the unweighted directed network with directed adjacency matrix given by M. $X_i(t)$ evolve according to* (1)*. Then the normalized vector $x_i(t) = X_i(t)/\sum_j X_j(t)$, converges to a stable fixed point $x := \lim_{t\to\infty} x(t)$,*

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

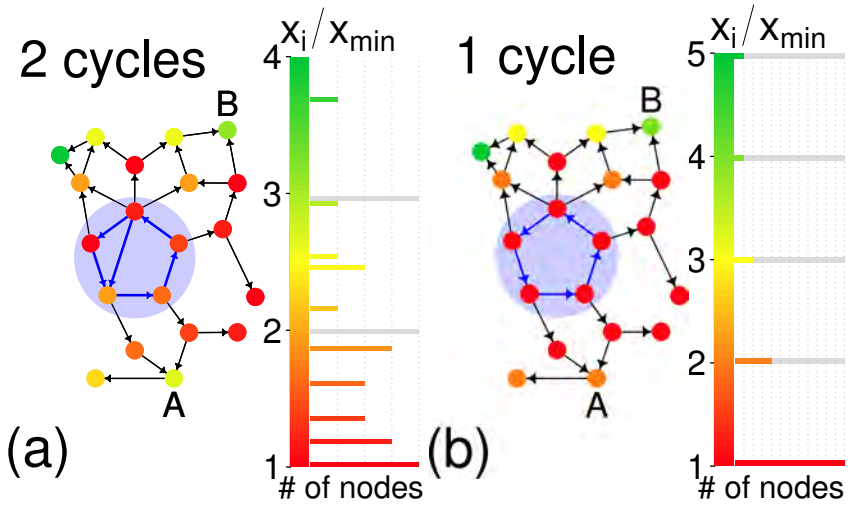


Fig. 1. Graphical demonstration of the quantization theorem. Directed networks M containing two cycles (a) and one cycle (b). Cycles are in the shaded area. The node color indicates the state, x_i, (value of the component of the state vector, x) in units of the minimal value $x_{\min}$. The histograms show the number of nodes in a given state. In (b) we see the quantization of states due to the presence of a single (!) remaining cycle. The occurrence of quantization at the last remaining cycle can be used as a precursor signal. The state $x_i/x_{\min}$ in the single cycle network (b) coincides with the number of directed paths from the cycle to node i. Node A can be reached through two paths from the cycle, while node B can be reached by four. For the multi-cycle case (a), the number of paths no-longer coincides with the states.

for which the following holds:
Eigenvector Quantization: *Suppose G contains only one single cycle. Then any component x_i can be expressed as*

$$x_i = n_i x_{\min} \,, \tag{2}$$

where $x_{\min}$ is the minimal non-zero component and n_i is a natural number. If there are no paths from cycle nodes, then $x_i = 0$.

The proof is based on the fact that the Perron-Frobenius eigenvalue $\lambda = 1$ if there is only one cycle. The graphical visualisation of the theorem is shown in Fig. 1. We compare the Perron-Frobenius eigenvectors for two network, one with two cycles and the other with only one cycle. The latter one exhibits the eigenvector quantization.

To put some flesh on the bare bones, we apply this result to one prominent example of autocatalytic models — Jain-Krishna model. Here the network evolves on the slow timescale, while the population $X_i(t)$ evolves according to Eq. (1) on much faster scale (i.e., it reaches the fixed point before the network is updated). In the updating procedure, one of the least populated nodes is removed and a new node is introduced. The node is connected with the existing nodes by assigning in-links to and out-links from the new species, both with the same probability $m/(N-1)$. Therefore m has the role of

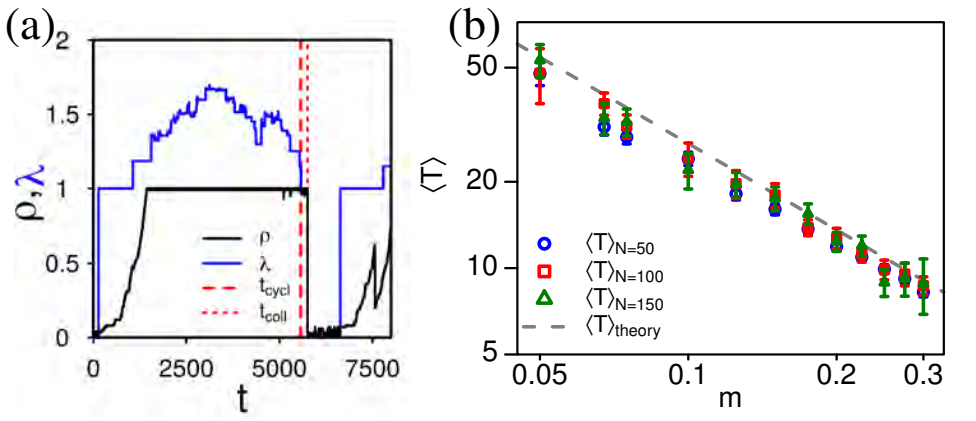


Fig. 2. (a) Sample run of the Jain-Krishna model. ρ is the fraction of populated nodes, λ is the largest eigenvalue. The time point t_{cycl} denotes the moment of entering the one-cycle phase (where the eigenvector quantization appears) and t_{coll} denotes the moment of collapse. We see that the collapse is preceded by the one-cycle phase. (b) Comparison between numerical simulations (symbols) and the theoretical prediction. (dashed line) of the expected time-to-collapse, $\langle T \rangle$, for various system sizes $N \in \{25, 100, 200\}$, and connectivities, m between 0.05 and 0.375. Simulation results follow the theoretical prediction closely, independent of system size and m.

average connectivity. The typical slow-scale dynamics is depicted in Fig. 2(a). The system starts in random phase, where no cycles are present. Once a cycle is created, the systems switches to ordered phase and the number of non-zero populated nodes starts to grow. Once it reaches its maximum, the system remains in the ordered phase until the collapse. It is possible to show that the collapse is preceded by the critical phase, where only one cycle remains. Therefore the eigenvector quantization serves as an early-warning signal that the collapse might come. Moreover, it is also possible to calculate the expected time to collapse which is (for large networks and reliable range of connectivity) simply equal to $T = \frac{e}{m}$, where e is the Euler's number. This prediction is in agreement with the simulations, as shown in Fig. 2.

Summary. We have presented an early-warning signal for a broad class of linear systems called eigenvector quantization. This measure does not require any structural information about network topology in order to anticipate the collapse of a system. We have shown that it can be successfully used in Jain-Krishna model, where we can also calculate the expected time to collapse. It can be also shown that this measure can be used in more general situations. Examples include non-linear systems as epidemic spreading in the SIS model, extension to other centrality measures as Katz centrality, or application to slightly weighted networks.

References

1. L. Horstmeyer, T. Pham Mihn, J. Korbel and S. Thurner, Predicting collapse of adaptive networked systems without knowing the network, in review.

Significant improvement of network robustness by enhancing loops through rewiring

Masaki Chujyo and Yukio Hayashi

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, JAPAN

1 Introduction

Many real networks have power-law degree distributions and are extremely vulnerable for malicious attacks [1]. However, it is discovered that onion-like networks with positive degree-degree correlations [2] have optimal robustness against the attacks [3][4]. To improve the robustness by enhancing the correlations, some methods based on edges swaps [5] or rewirings [3] have been proposed. In addition, an efficient algorithm proposed by Z.-X. Wu and P. Holme generates nearly optimal robustness [6].

In this paper, for improving the robustness of connectivity, we focus on enhancing loops measured by feedback vertex set (FVS) which is a subset of vertices that removal makes the graph cycle-free, because network dismantling and decycling problems can be considered as asymptotically equivalent [7]. Furthermore, it has been pointed out that the robustness is stronger as a larger fraction of FVS in incrementally growing onion-like networks [8]. Thus, enhancing loops on a network is crucial for improving network robustness. The purpose of this study is to clarify a deeper relation between robustness and loops than correlations. In previous works, robustness is usually discussed in a relation to degree-degree correlations. We propose new methods to enhance loops and numerically show a significant improvement of robustness by enhancing loops.

Followings are the background. Although the minimum FVS problem is intractable called as NP-hard, there exists an efficient approximation algorithm by applying belief propagation (BP) based on a cavity method in statistical physics [9]. It calculates marginal probability q_i^0 for the state 0 of node i, which denotes the candidate probability of belonging to FVS. In another related topic, increasing the number of spanning tree by rewiring is corresponded to enhancing loops. The number of spanning tree represent varieties of loops in a network, since each edge that does not belong to a spanning tree is one-to-one corresponding to a loop, whose set consists of a linearly independent basis called fundamental cycles [10]. In other words, any cycle on a network can be represented by a linear combination of the fundamental cycles.

2 New type of rewiring to enhance loops

We introduce new rewiring methods based on edge rewiring with or without degree-preserving in order to enhance loops. Increasing the size of FVS means enhancing loops. When low q^0 nodes are connected, it is expected that the size of FVS increases, because they are not concerned with any loops and make new loops by the connections.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

To keep the degree in increasing the size of FVS, we remove edges between low and high q^0 nodes, and add two edges: an edge between low q^0 nodes and an edge between high q^0 nodes. The degree-preserving method is summarised as follows.
Step 1.) Select two nodes i and j which are disconnected and have the highest $q_i^0 + q_j^0$.
Step 2.) Select a node l which is the lowest q_l^0 node in the neighbor nodes of node j. Similarly, select a node k which is the lowest q_k^0 in the neighbor nodes of node i and unconnected to node l.
Step 3.) Add edges (i, j) and (k, l), and remove edges (j, l) and (i, k).
Instead of degree-preserving, in another non-preserving method [1] , the differences are
Step 1.) Remove an edge (i, j) which have the highest $q_i^0 + q_j^0$.
Step 2.) Add an edge (k, l) which have the lowest $q_k^0 + q_l^0$.
For comparison, we also consider other methods in which q^0 is replaced with degree.

3 Deeper relation of robustness and loops than correlations

We numerically investigate the improvement of robustness for our proposed method in comparing with other conventional methods [6] [10]. We apply them to 10 real networks. Figure 1 shows typical results for the robustness against hub attacks [3] by measuring the size of the giant component, the number of nodes in the FVS [9] and assortativity for the correlations [2] versus increasing the number of rewiring. In particular, degree-non-preserving methods improve robustness more than degree-preserving methods. The result suggests that degree-preserving has a restriction on improving robustness. In comparison with the rewiring on OpenFlights at #Rewire=8000, our degree-non-preserving method is twice as robust as the best in degree-preserving methods. In Fig 1, the ordering of lines in robustness is almost the same as in FVS. However, they are greatly different from the ordering in assortativity. Therefore, the robustness is related to the size of FVS rather than assortativity focused conventionally.

Summary. We propose rewiring methods to enhance loops and obtain the result with significant improvement of robustness by enhancing loops. Moreover, our result suggests that loops are more essential for the robustness than the degree correlations.

References

1. R. Albert, H. Jeong, and A.-L. Barabási, Nature 406(6794), 378 (2000)
2. M.E.J. Newman, Phys. Rev. Lett. 89(20), 208701 (2002)
3. C.M. Schneider et al., PNAS 108(10), 3838-3841 (2011)
4. T. Tanizawa, S. Havlin and H.E. Stanley, Phys. Rev. E 85(4), 046109 (2012)
5. V.H. Louzada et al., J. Complex Net. 1(2), 150-159 (2013)
6. Z.-X. Wu and P. Holme, Phys. Rev. E 84(2), 026106 (2011)
7. A. Braunstein et al, PNAS 113(44), 12368-12373 (2016)
8. Y. Hayashi and N. Uchiyama, Sci. Rep. 8(1), 1-13 (2018)
9. H.-J. Zhou, EPJB 86(11), 455 (2013)
10. H. Chan and L. Akoglu, Data Min. Knowl. Discov. 30(5), 1395-1425 (2016)

Acknowledgements: This research is supported in part by JSPS KAKENHI Grant Number JP.17H01729.

[1] They are the best selections for the robustness in operations several combinations for the highest or the lowest nodes and the ordering of add and remove.

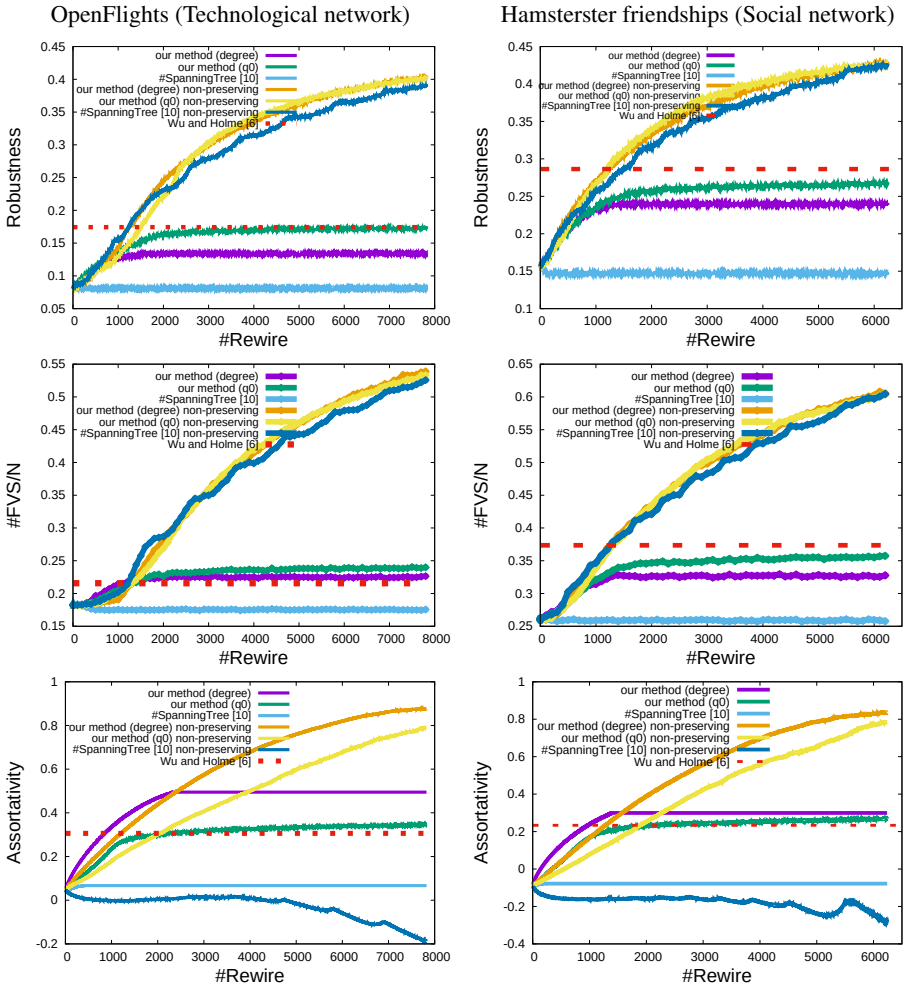


Fig. 1. Comparing the results for our proposed and the conventional rewiring methods in real networks. The red dotted line indicates the conventional best case for the robustness in each network. The large difference between green and yellow lines for the robustness is caused by whether or not degree-preserving. In the ordering, the green line is higher than the violet line in both the robustness and the size of FVS but lower in the assortativity. The data can be obtained from OpenFlights: http://konect.uni-koblenz.de/networks/opsahl-openflights, Hamsterster friendships: http://konect.uni-koblenz.de/networks/petster-friendships-hamster.

Identifying a crucial role for robustness and spreading in complex network

Fuxuan Liao and Yukio Hayashi

Japan Advanced Institue of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan

1 Introduction

Although it has been believed for a long time that a high degree node (hub) is the best selection for effective spreading and malicious attack, exciting research of network science has recently unveil that k-core is more important than hub (node) in some sense. The most efficient spreaders are located within the core of network as identified by the k-shell decomposition analysis [1]. For example, if a high degree node in a dangling sub-tree is not necessarily effective for spreading, this node belongs to 1-shell. The spreading from such node in a dangling sub-tree quickly disappears, while the spreading can persist in the core. Thus, in order to find out influence nodes for spreading, a node with high k-shell index is more useful than high degree nodes [1]. Here, k-core is defined as a connected remaining part by recursively removing nodes whose degrees are less than k [2]. The k-shell is the part by deleting (k+1)-core from k-core.

On the other hand, from the asymptotic equivalence of dismantling and decycling problems at infinite graphs in a large class of random networks with light-tailed degree distribution [3], the strong robustness is related to increasing the size of feedback vertex set (FVS) which is necessary to form loops [4]. The existence of many loops may be crucial to maintain the connectivity of network within a finite size. Dismantling (or decycling) problem is to find the minimum set of nodes whose removal yields a graph with the largest connected cluster whose size is at most a constant (or a graph without loops) [3]. Moreover, based on the definition of influencer in Collective Influence (CI), when the minimum set of important nodes for spreading is removed, the propagation is stops; At that time, the network becomes a tree without loops at the critical just before destroying the connectivity [5]. Thus, loops are also related to influencer. From the above viewpoints, it is necessary for improving the robustness to exist many loops in a network, stronger robustness have larger size of FVS. In this paper, we find that not only degree and k-shell are important for network to find out the crucial nodes for spreading and robustness, but also FVS plays the same role. More precisely, our goal is to clarify the relations between FVS and k-shell.

2 Fraction of FVS nodes in k-shells

In order to clarify the relation between FVS and k-shell, we need to get the fraction of FVS nodes in k-shells. The minimum size belongs to NP-hard problem, there is no efficient algorithm to obtain the exact solution. Thus, based on the cavity method in

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

statistical physics, we apply efficient belief propagation (BP) algorithm proposed by H-J Zhou [6] to calculate candidate nodes of the FVS. Then, we divide the set of nodes to make the k-shell decomposition into different shells. Finally, the fraction of FVS nodes in each k-shell is calculated.

We investigate the relation between FVS and k-shell for the data of 28 real scale-free networks [7] gathered from different fields. Figure 1 shows typical results in three examples. Here, N, $\mathscr{F}$ and K_s denote the size (the number of nodes included in each subset) of the whole network, FVS, and k-shells, respectively. As shown by red line in Fig. 1, the fraction $\frac{|\mathscr{F} \bigcap K_s|}{|K_s|}$ of FVS in k-shell is increasing as the k-shell index is larger. In other words, the inner core contains a higher fraction of FVS. As shown by blue lines in Fig. 1, the fraction $\frac{|K_s|}{|N|}$ of k-shell in the network size N is decreasing as the k-shell index is larger. Therefore, the inner core is consisted of a few nodes. In blue line, $|K_s|$ is smaller as larger k-shell index, the intersection to FVS also becomes small in the whole FVS, this decreasing is shown by green line. However, in some case, depending on the zig-zag shapes in red and blue lines, there are peaks in green lines.

In Fig. 2, in order to make the results more intuitive, we visualize the k-shell decomposition. Colored nodes except gray ones in each circle represent a node in FVS. The red lines show the connections of FVS nodes in a same shell, yellow lines show the connections between nodes in FVS included in different shells. As shown small rings in the Fig. 2, the large k-shell indexes have the high fractions of FVS, even if each $|\mathscr{F} \bigcap K_s|$ is small.

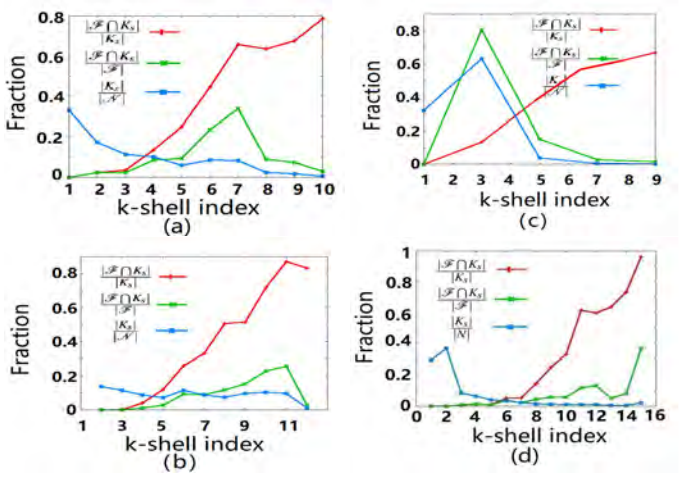


Fig. 1. Results for real networks in four different domains. (a) biological network of yeasts with N = 2224, # of FVS = 363, maximum k-shell index k = 10 (b) technological US power grid with N = 4941, # of FVS = 516, maximum k-shell index k = 5 (c) social email networks with N = 1133, # of FVS = 370, maximum k-shell index k = 12 (d) Japanese language networks with N = 2698, # of FVS = 136, maximum k-shell index k = 15.

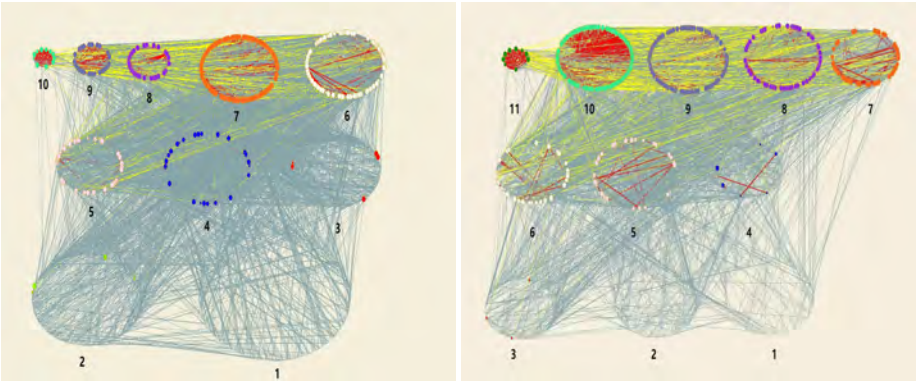


Fig. 2. Visualization by Pajek as typical result. From left to right, biological network of yeasts with N = 2224, # of FVS = 363, maximum k-shell index k = 10, technological US power grid with N = 4941, # of FVS = 516, maximum k-shell index k = 12. The numbers represent k-shell indexes.

Summary. Our research is an extension to investigate the relation between degree and k-shell [1]. We suggest that FVS and k-shell play a crucial role for robustness and spreading. We will elucidate that the nodes in FVS become important not only for robustness but also for efficient spreading. In addition, to find FVS is a NP-hard problem, while the k-shell decomposition is a P-problem; It is expected that the gap in computational effort leads to develop a new direction by solving the P-problem to estimate the FVS.

Acknowledgements

This research is supported in part by ISPS KAKENHI Great Number JP.17H01729.

References

1. M. Kitsak et al. Nature Physics 6, 888-893 (2010).
2. J.I.A. Hamelin. Advances In HIPS 18, Canada (2006).
3. Y. Hayashi, N. Uchiyama. Scientific Reports 8, 11241 (2018)
4. R.M. Karp, E. Miller et al. (eds), pp.85-103, NY Plenum Press (1972).
5. F. Morone, H.A. Makse. N 65, vol 524, (2015).
6. H-J Zhou. EPJB, DOI: 10.1088, 1742-5468 (2016).
7. F. Radicchi. Phys. Rev. E 91, 010801(R) (2015)
8. A. Braunstein et al. PNAS 113(44), 12368-12373 (2016).
9. S. Mugisha, H-J Zhou. Phys. Rev. E 94, 012305 (2016).

Network clustering-based design of controllable and observable dynamical systems with small relative degree

Dániel Leitold[1,2], Ágnes Vathy-Fogarassy[1,2], and János Abonyi[2]

[1] Department of Computer Science and Systems Technology
University of Pannonia Egyetem st. 10., Veszprém, H-8200, Hungary
[2] MTA-PE Lendület Complex Systems Monitoring Research Group
University of Pannonia Egyetem st. 10., Veszprém, H-8200, Hungary
janos@abonyilab.com,
WWW home page: https://www.abonyilab.com

1 Introduction

The application of network science in the field of dynamical systems enabled to interpret the dynamical properties of the systems in the form of network representation [1, 2]. During the application of the approach, it can be recognised that the assigned driver and sensor nodes are only a small proportion of the state variables (nodes) and the results are not applicable/realistic [3]. To reduce the resulted high relative degree that makes the design of the controller difficult and results in sluggish control performance, additional actuators and sensors are necessary to assign to the existing configurations [4].

In the network interpretation of any linearised or linear system, assuming all node is driver node, the relative degree r_{ij} can be defined as the length of shortest path from sensor node y_i to driver node u_j, as it can be seen in Figure 1. The relative degree of output u_j is the minimum of these lengths, $r_i = \min_i r_{ij}$. The relative degree of the system is the maximum of all r_i, $r = \max_i r_i$. As a result, the network is segmented according to the shortest paths between the outputs and the drivers closest to them.

To decrease relative degree, we used four methodologies that ensure observability beside the minimisation the relative degree. Two of these methods utilises simple heuristics [5]. The first one operates with closeness and betweenness centrality measures while the second one creates and solves a set covering problem from the sensor placement problem. The second two methods combine the meta-heuristic simulated annealing optimisation with clustering [4].

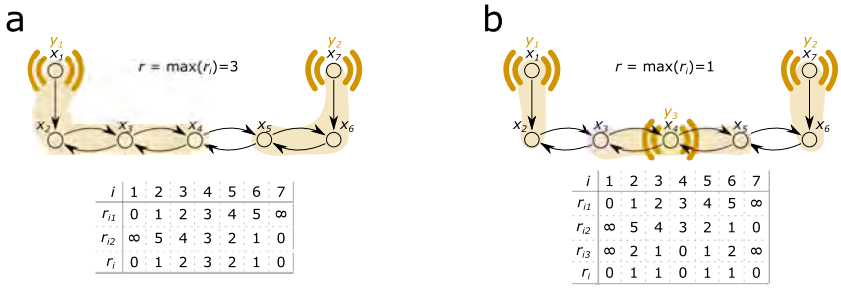
2 Results

Beside the reduced relative degree, the generation of a balanced monitoring is of interest, so the following cost function was applied through the determination of the set of sensor nodes S:

$$\min_S cost(G,S,\beta) = \beta \max_{i=1}^{K} r_i + (1-\beta)\frac{\sum_{i=1}^{K} r_i}{N}, \tag{1}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



i	1	2	3	4	5	6	7
r_{i1}	0	1	2	3	4	5	∞
r_{i2}	∞	5	4	3	2	1	0
r_i	0	1	2	3	2	1	0

i	1	2	3	4	5	6	7
r_{i1}	0	1	2	3	4	5	∞
r_{i2}	∞	5	4	3	2	1	0
r_{i3}	∞	2	1	0	1	2	∞
r_i	0	1	1	0	1	1	0

Fig. 1. Illustration of the concept of relative degrees and how it segments a simple toy network with **(a)** two or **(b)** three sensors are assigned. We assume that all node is driver node. It is visible that one additional sensor decreases the relative degree of the system significantly.

where G denotes the network representation of the system, S stands for the set of assigned sensors, parameter $\beta = [0,1]$ weights between the balance-related average and the maximum of the relative degree of the system, N is the number of nodes (state variables) and $K = |S|$ is stand for the number of sensors assigned to the system.

Four methods are presented to handle this problem:

- *centrality measures-based method* [5]: The first approach utilises the closeness centrality and the betweenness centrality measures to determine the position of the additional sensor nodes. Firstly, the network is segmented to the minimal configuration, S, generated by maximum matching and ensures structural observability. Then for each segment, based on the closeness and betweenness centrality the node with the highest centrality became a new sensor node.
- *set covering-based method* [5]: The second approach creates a set-covering problem as follows: for each node (state variable) the node set is generated that can be reached in a maximum r_{max} length path. Then the nodes are covered by the sets of sensors of minimal configuration, S, removed from all sets. The remaining sets create a set-covering problem. Solving the well-known problem, the resulted sets determine the nodes should be observed.
- *modified Clustering Large Applications based on Simulated Annealing (mCLASA)* [4]: The mCLASA methods extend the minimal configuration, S, with additional sensors assigned randomly to the network. Simulate annealing granted that the result converges to an optimum.
- *Geodesic Distance-based Fuzzy c-Medoid Clustering with Simulated Annealing (GDFCMSA)* [4]: The GDFCMSA method works similarly to mCLASA, but for determining the location of the additional sensors, it uses a fuzzy c-medoid clustering.

The method is applied to the sensor placement of Heat Exchanger Networks (HENs) [4] having 10-100 units (nodes). The centrality measures-based method assigns more sensor to the system than other methods, thus it segments more the network than other method. It favours to the relative degree, as smaller segments mean shorter paths. The other three methods have almost the same performance. To highlight the differences, we examined how many times a method gives better or equal result, and how many

times were a member of the Pareto frontier in the analysis of more than 600 HEN [6] (Table 1). The comparison is based on K with given r_{max}. The Pareto frontiers were based on K and r (2-dimensional), and based on K, r and *cost* (3-dimensional). The results show that centrality measures-based method is the worst method if the goal is to minimise the number of sensors. The small differences between the set covering-based, mCLASA and GDFCMSA algorithms can be pointed out, as in the same order the result is getting better and better. The results can be generated for the controllability problem as well, as it is the mathematical dual of observability.

Table 1. Comparison of the proposed methods as how many times a method (row) gives better and equal solution than other (column), and how many times was the method a member of the Pareto frontier.

	CentMeas	**SetCov**	**mCLASA**	**GDFCMSA**	**CentMeas**	**SetCov**	**mCLASA**	**GDFCMSA**	**2-dimensional**	**3-dimensional**
	Better				Equal				Pareto	
CentMeas	0	0	2	1	0	116	109	107	622	631
SetCov	523	0	69	45	116	0	527	533	605	638
mCLASA	528	43	0	17	109	527	0	556	614	638
GDFCMSA	531	61	66	0	107	533	556	0	612	639

References

1. Lin, Ching-Tai. Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201-208, 1974.
2. Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167, 2011.
3. Dániel Leitold, Ágnes Vathy-Fogarassy, and János Abonyi. Controllability and observability in complex networks–the effect of connection types. *Scientific reports*, 7:151, 2017.
4. Daniel Leitold, Agnes Vathy-Fogarassy, and Janos Abonyi. Network distance-based simulated annealing and fuzzy clustering for sensor placement ensuring observability and minimal relative degree. *Sensors*, 18(9):3096, 2018.
5. Dániel Leitold, Ágnes Vathy-Fogarassy, and János Abonyi. Design-oriented structural controllability and observability analysis of heat exchanger networks. *Chemical Engineering Transactions*, 70:595–600, 2018.
6. Dániel Leitold, Ágnes Vathy-Fogarassy, and János Abonyi. Evaluation of the complexity, controllability and observability of heat exchanger networks based on structural analysis of network representations. *Energies*, 12(3):513, 2019.

No-exclaves percolation: Uncovering hidden impact of failures in complex systems

Sang-Hwan Gwak, Eunkyu Park and K.-I. Goh

Department of Physics, Korea University, Seoul 02841, Korea
kgoh@korea.ac.kr,

The subject of network robustness has been one of the major topics of complex network theory since its inception. To shed a new light on this important subject, we introduce a new cluster called an '*exclave*' cluster and study its percolation properties regarding how it is related to collapses in networks. The exclaves cluster in our model is a modification of the No-enclaves percolation(NEP) model proposed by Sheinman *et al* defined on $2D$ Euclidean lattices to make it applicable to networks. When some nodes in a network fail or get removed, there can emerge sets of connected non-failed nodes that are completely surrounded by the failed nodes called the exclaves. The exclaves are isolated from other functioning parts of the network, thereby becoming effectively non-functional. This process defines a new class of cluster of non-functional nodes (the no-exclaves or NExP cluster), formed by the connected union of failed nodes and exclaves. Thus, the NExP clusters account for the hidden impact of network failure that had never been investigated as yet. Our main aim in this presentation is to report essential understanding gained by the new measure and to offer a new perspective of study of network robustness.

First, we will show the mechanism for forming NExP clusters through empirical networks. We introduce the behavior of NExP clusters in urban road networks as our model application. In application of urban road networks, NExP cluster corresponds to the isolated cluster subject to the road closure due to traffic malconditions in the road network. Moreover, we examine the empirical network using finite size scaling method through different road length to interpret the critical phenomena.

Second, by a novel generating function-based analytic theory on random networks as well as extensive Monte Carlo simulations, we uncover that the NExP display profoundly different results from those of ordinary percolation (OP). We found that the NExP displays two distinct transitions: One between non-percolating to percolating phase at q_c, and the other between the partially-percolating to fully-percolating phase at q^*. The percolation transition at q_c occurs at much lower failure probability than corresponding OP and the size of NExP giant cluster is significantly larger than that of OP, both manifestly demonstrating the hidden impact. We measure critical exponents and found them consistent with mean-field class. Our study discloses hidden indirect damage additional to damage from direct attacks, and thus suggests a new useful way for finding non-functioning areas in complex systems.

References

1. Sheinman, M., et al. "Anomalous discontinuity at the percolation critical point of active gels." Phys. Rev. Lett 114.9 (2015): 098104.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

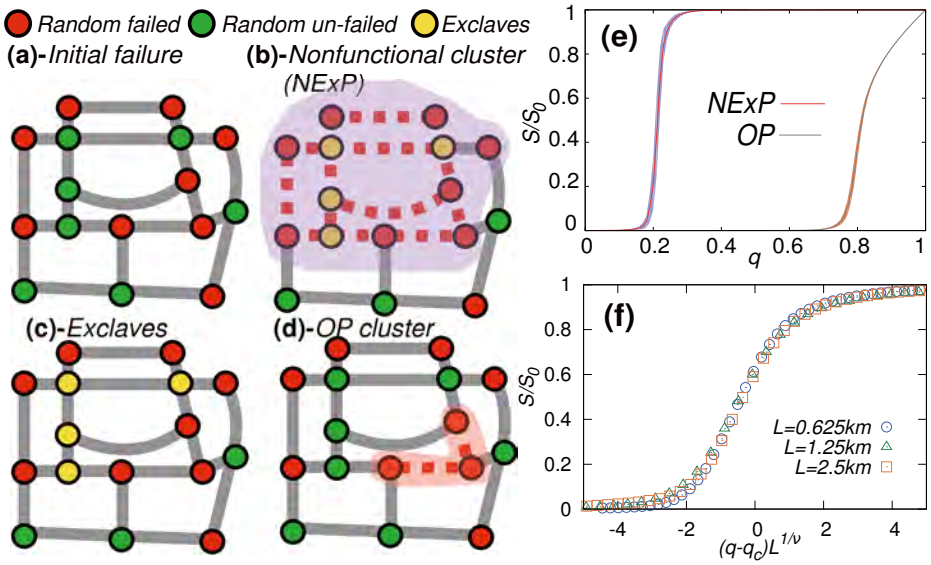


Fig. 1. (a, b, c, d) are schematic illustration of NExP formation process on schematic diagram of road networks. Marked by red nodes are failed node and green nodes are un-failed node with probability q, $1-q$ respectively. Yellow nodes are exclaves nodes (or clusters) which are completely surrounded by failed nodes, thus NExP cluster is union of exclaves and failed cluster highlighted with purplish shades shown in (b). The area that cannot actually function as a road is a union NExP cluster, not a reddish OP cluster. (e) is effective giant connected component S/S_0 vs. failed prob. q of NExP and OP on London road network. (f) Finite-size scaling using data collapse of simulation results with different road length. Those are well collapsed onto single line with critical exponents $\beta = 0, \gamma = 2\nu = 8/3$. In the collapse system, exclave clusters not only lead to rapid collapse but also lead to changes of transition types.

K-selective percolation on complex networks

Jung-Ho Kim and K.-I. Goh

Department of Physics, Korea University, Seoul 02841, Korea,
kgoh@korea.ac.kr

1 Introduction

Percolation theory provide the theoretical foundation for how complex networks react to intentional attacks and random errors [1]. Until now, many researches have been conducted on the intentional attacks to high centrality nodes [2]. However, nodes with high centrality are not the only components in complex networks. Rather, intermediate and low centrality nodes occupy the most fraction of complex networks. In the case of degree centrality, attacks on low centrality nodes can be analyzed using k-core percolation model [3]. However, attacks on intermediate centrality nodes has not been studied. We made a new percolation model, the K-selective percolation, to observe how complex networks respond to attacks on nodes with intermediate degree.

2 Model

The K-selective percolation has following simple rules [Fig.1.(a)]. Select a random node and delete it if the selected node has degree K. This process continues until there are no more nodes having degree K in the complex network. We applied the K-selective percolation rule after bond-deletion with probability q, and we use q as control parameter. The order parameter (M) is defined as the probability that the randomly-chosen node belongs to the giant cluster.

3 Results

We applied the K-selective percolation to the scale-free network ($k_{min} = 5$, $k_{max} = 100$, $2.5 \leq \gamma \leq 4$) made by configuration model and Erdős-Rényi network. We derived numerical solutions using generating function method and verified them by extensive Monte Carlo simulation. The most signature feature of K-selective percolation is the existence of fragile valleys that are vulnerable points of complex systems. On scale-free network, fragile valleys appear near $q = 0.3$ [Fig.1.(b)]. Hybrid phase transitions can occur when entering and leaving fragile valleys as q increases. After systems leave the valley and cross the hill, continuous phase transitions are observed near $q = 0.8$ and the giant cluster disappears [Fig.1.(b)]. We obtained qualitatively similar results on Erdős-Rényi networks. We studied the critical behavior of phase transitions using finite-size scaling theory on Erdős-Rényi network. We obtained similar critical exponents with k-core percolation [4] on hybrid phase transition. In case of continuous phase transition, we conclude that it belongs to the same universality class with ordinary percolation.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

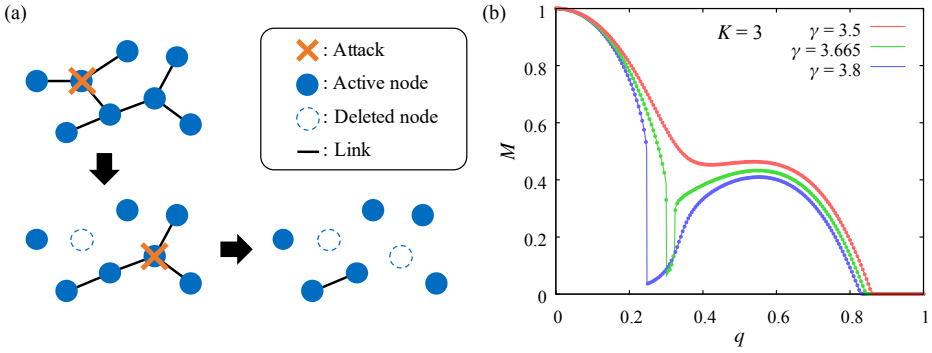


Fig. 1. (a) A schematic example for 3-selective percolation on simple network. (b) Phase diagram of 3-selective percolation on scale-free network with $k_{min} = 5$, $k_{max} = 100$. Lines are numerical solution and points are Monte Carlo simulation with $N = 10^7$.

References

1. Albert, R., Jeong, H., Barabsi, A. L.: Error and attack tolerance of complex networks. Nature 406, 378382 (2000).
2. Holme, P., Kim, B. J., Yoon, C. N., Han, S. K.: Attack vulnerability of complex networks. Phys. Rev. E 65, 056109 (May 2002)
3. Dorogovtsev, S. N., Goltsev, A. V., Mendes, J. F. F.: k-Core Organization of Complex Networks. Phys. Rev. L 96, 040601 (Feb 2006).
4. Lee, D., Jo, M., Kahng, B.: Critical behavior of k-core percolation: Numerical studies. Phys. Rev. E 94, 062307 (Dec 2016).

Coupling transport and supply-chain networks to evaluate the indirect impact of disasters — application to the United Republic of Tanzania

Celian Colon[1], Stephane Hallegatte[2], and Julie Rozenberg[2]

[1] International Institute of Applied System Analysis, Laxenburg, Austria
celian.colon@polytechnique.edu
[2] The World Bank, Washington DC, USA

1 Introduction

Transport systems play a pivotal role in connecting multiple components of an economy. They enable firms to source inputs from distant suppliers and deliver their outputs to customers. But transport networks are vulnerable to natural disasters, such as floods, landslides, or earthquakes. Roads may become impassable, forcing trucks to take a longer, sometimes busier, itinerary. Disasters can even paralyze an entire transport node, such as a port or an airport, inducing severe delays.

Existing methods quantify the direct impacts of transport disruption [1][2]. By modeling the full network, they estimate how much trade and people flows are blocked, and evaluate the cost of rerouting flows. Criticality maps can be drawn to prioritize interventions. But what does these rerouted or blocked flows mean for the economy? How do they affect the different productive sectors? The indirect economic impacts are missing.

Input–output models evaluate such indirect economic impacts, but at a very aggregated levels [3][4]. Localized shocks are translated into a reduction in sector-level production, which is then propagated using the input–output interlinkages. Such degree of aggregation is likely to underestimate the indirect loss triggered by supply-chain disruptions [5].

In this project, we formulate a model that estimates the indirect economic impacts of transport disruptions. We apply the model to the United Republic of Tanzania, whose fast growing economy is vulnerable to climate-related disasters, especially floods.

2 Method: downscaled supply chains embedded on roads

We model the road network of the United Republic of Tanzania, which carries more than 99.5% of the trade flows. Using an agent-based modeling framework, we populate the road network with firms and households.

Using a variety of data—including input–output tables, business survey, population and land cover data—we spatially disaggregate sectoral production and consumption. Production is assigned to firms and consumption to households. Firms have a Leontief production function calibrated with input–ouptut data. They hold inventories for each type of inputs, calibrated with survey data. In the absence of available supply-chain

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

data, we reconstruct the supply chain network, based on sectors, distance, and firm size. Firms dynamically order, produce, and deliver goods to their clients. In the absence of perturbations, the model is at a steady state.

When a node or a link of the transport network is disrupted, we model two mechanisms to capture the indirect economic losses.

- If an alternative pathway exists, supply chain flows are rerouted, leading to higher transport costs. Suppliers transfer these higher costs onto their clients by increasing their selling price. In turn, clients transfer such increase in input costs to their own clients. Price increases propagate down to households, which have to spend more for the same basket of goods.
- If an alternative pathway does not exist, products that were supposed to be delivered are hold at the producers premises. This situation, if prolonged, may lead to a shortage of inputs for the producers' clients. At the end of the supply chains, households may have to decrease their consumption.

Both mechanisms allow us to compute the indirect losses caused a disaster at the door of the final consumers. These losses capture all the impacts that cascade along the supply chains.

We also incorporate imports, exports, and transit flows. To that end, agents representing countries that trade with the United Republic of Tanzania are modeled.

3 Results

These indirect losses generally affect people that are not directly hit by disasters. Their intensity nonlinearly increases with the duration of the initial disruption; see Fig. 1. Supply chains generate interdependencies that amplify disruptions for nonprimary products, such as processed food and manufacturing products.

We identify bottlenecks in the network. But their criticality depends on the supply chain we are looking at. For instance, some infrastructures are critical to some agents, say international buyers, but of little use to others. Investment priorities vary with policy objectives, e.g., support health services, improve food security, promote trade competitiveness.

Resilience-enhancing strategies can act on the supply side of transportation, by improving the quality of targeted infrastructure, developing alternative corridors, building capacity to accelerate post-disaster recovery. On the other hand, policies could also support coping mechanisms within supply chains, such as sourcing and inventory strategies. Our results help articulate these different policies and adapt them to specific contexts.

(1)

Summary.

Fig. 1. Supply-chain impacts on households triggered by disruptions of one to four weeks. Each bar represents a distribution of impacts obtained by disrupting the 300 most critical transport nodes. The filled rectangle indicates the interquartile interval, the solid horizontal lines indicates the median, the dash horizontal line indicates the mean. Mean values are joined together with a black curve. The vertical line extends to the minimum and maximum of the distributions; when the maximum lies outside the plotting area, the upper part of the vertical line is not capped.

References

1. Rozenberg J, Briceno-Garmendia C, Lu X, Bonzanigo L, Moroz H (2017) Improving the Resilience of Perus Road Network to Climate Events. Policy Research Working Paper 8013, The World Bank Group, Washington DC, USA.
2. Espinet X, Rozenberg J, Singh Rao K, Ogita S (2018) Piloting the Use of Network Analysis and Decision-Making under Uncertainty in Transport Operations Preparation and Appraisal of a Rural Roads Project in Mozambique under Changing Flood Risk and Other Deep Uncertainties. Policy Research Working Paper 8490, The World Bank Group, Washington DC, USA.
3. Haimes YY, Jiang P (2001) Leontief-Based Model of Risk in Complex Interconnected Infrastructures. Journal of Infrastructure Systems 7 (1), 112.
4. Okuyama Y (2004) Modelling spatial economic impacts of an earthquake: inputoutput approaches. Disaster Prevention and Management: An International Journal 13 (4), 297306.
5. Henriet F, Hallegatte S, Tabourier L (2012) Firm-network characteristics and economic robustness to natural disasters, Journal of Economic Dynamics and Control 36, 150167.

Control of core-periphery networks under sparse feedback controllers

Ilias Mitrai, Wentao Tang, and Prodromos Daoutidis

University of Minnesota, Minneapolis MN 55455, U.S.A.
daout001@umn.edu

1 Introduction

Controllability has been studied by network scientists using classical control-theoretic concepts such as structural controllability and control energy [1]. Many works, *e.g.*, [2], have attempted to reveal some fundamental relations between topological features and controllability of networks. However, controllability is an open-loop control property which does not account for the specific control mechanism. This inevitably neglects two important aspects of real-world control of networks. First, control actions are executed by *feedback* controllers. Second, the information exchange between the nodes of the network and its controller brings an additional cost, which implies that controller *sparsity* is important.

In our recent research, we adopted the sparsity-promoting control framework [3] to evaluate network control cost based on both control performance and sparsity of the feedback controller, with the optimal controller obtained by minimizing the combined performance and sparsity costs. These works [4, 5] aim at understanding the role of *community structures*, namely blocks of nodes that are densely interconnected inside but weakly coupled between, in improving network control, when a significant cost is associated with the number of feedback channels to promote the sparsity of the feedback controller. Furthermore, we have shown [6,7] that the common existence of community structures in networks, such as those in the biological world, can be interpreted as the result of an evolutionary process driven by a combination of control performance and sparsity.

Core-periphery (CP) structure is another type of common topological feature in biomolecular, ecological, social and traffic networks [8]. It refers to a dichotomy between a "core" (C) part of the network, whose nodes have stronger propensity to be connected to other nodes, and the remaining "periphery" (P) part. A variety of algorithms to detect CP in networks have been reviewed in [9]. In this work, we investigate the *effect of CP structure on network control cost* in terms of the total cost accounting for control performance and feedback sparsity.

2 Results

We create an ensemble of 3000 networks using a stochastic block model (SBM) [10]. The probability of creating an edge between C-C, C-P and P-P node pairs are set as θ_C^2, $\theta_C\theta_P$ and θ_P^2, respectively, where $\theta_C, \theta_P > 0$ stand for the probabilities of attaching a

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

half-edge to a C node and P node, respectively. With the total number of nodes ($N = 100$) and expected total number of edges ($E = 400$) fixed, the size of the core (N_C) and the half-edge probability of core nodes (θ_C) are the two degrees-of-freedom in network generation. A sample of the generated networks are shown in Fig. 1.

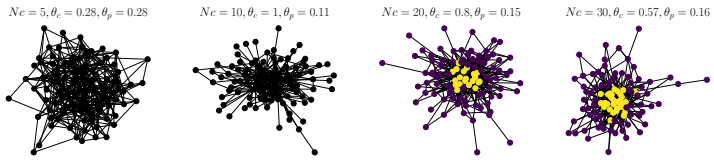


Fig. 1. Networks with different CP structure.

The generated networks are then associated with a Laplacian dynamics $\dot{x} = -Lx + d + u$, where L is the Laplacian matrix, and states x, disturbances d, and control inputs u are vectors with N components corresponding to the nodes. For a feedback controller $u = -Fx$ where F is the feedback gain matrix, the cost related to the controller performance $J(F)$ is the L_2-norm of the closed-loop system $d \to z = (y,u)$, and the cost for sparsity is the cost for each feedback channel γ multiplied by the number of nonzero entries in F (card(F)). The total control cost is then given by the minimized cost $J(F) + \gamma$card(F) under the optimal controller gain F.

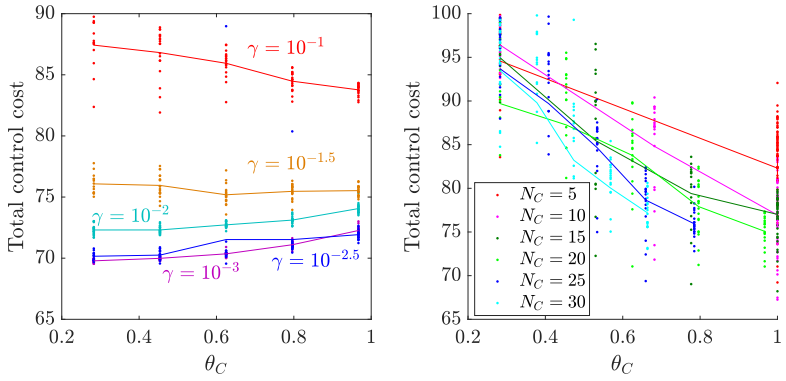


Fig. 2. Variation of the total control cost with different θ_C under (left) different γ and fixed $N_C = 20$, and (right) different N_C and fixed $\gamma = 10^{-1}$.

The total control cost of the networks are plotted against θ_C in Fig. 2. While at lower values of γ (when feedback sparsity is not accounted for), larger θ_C (higher extent of CP structure) increases the total control since the network is less well mixed. However, when the cost of feedback channels γ increases, networks with clearer CP structure gradually gain advantage in lowering the total control cost. Also, larger cores lead to lower control cost under fixed core interconnection density θ_C. But the core size is limited by the total edge number, and hence the lowest average cost is reached at an optimal

N_C of about 20. The adjacency matrices and the corresponding optimal controllers are visualized in Fig. 3. It can be seen that with clearer CP structures, the feedback controllers are more easily sparsified, with the feedback channels concentrated inside and around the core.

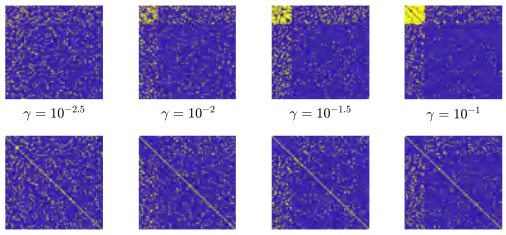


Fig. 3. Adjacency matrices and optimal feedback gains ($\gamma = 10^{-1}$) of networks with different CP structures. Yellow and blue pixels stand for nonzero and zero entries, respectively.

Summary. In this work, we have investigated the control of Laplacian networks with CP structure using a sparsity-promoting framework. It is found that under significant cost of feedback channels, CP networks have lower control cost compared to non-CP ones by adopting sparser controllers where the feedback channels are concentrated inside and around the cores.

References

1. Liu, Y.-Y., Barabási, A.L.: Control principles of complex systems. Rev. Mod. Phys. 88(3), 035006 (2016)
2. Pósfai, M., Liu, Y.-Y., Slotine, J. J., Barabási, A. L.: Effect of correlations on network controllability. Sci. Rep. 3, 1067 (2013)
3. Lin, F., Fardad, M., Jovanović, M. R.: Design of optimal sparse feedback gains via the alternating direction method of multipliers. IEEE Trans. Autom. Control, 58(9), 2426–2431 (2013)
4. Tang, W., Daoutidis, P.: The role of community structures in sparse feedback control. Am. Control Conf., 1790–1795 (2018, June)
5. Constantino, P. H., Tang, W., Daoutidis, P.: Topology effects on sparse control of complex networks with Laplacian dynamics. Sci. Rep. 9, 9034 (2019)
6. Constantino P. H., Daoutidis P. A control perspective on the evolution of biological modularity. 5th IFAC Conf. Intell. Control Autom. Sci., (2019, August)
7. Tang W., Constantino P. H., Daoutidis P.: Optimal sparse network topology under sparse control in Laplacian networks. 8th IFAC Workshop Distribut. Estim. Control Netw. Syst., (2019, September).
8. Csermely, P., London, A., Wu, L. Y., Uzzi, B.: Structure and dynamics of core/periphery networks. J. Complex Netw. 1(2), 93–123 (2013)
9. Rombach, P., Porter, M. A., Fowler, J. H., Mucha, P. J.: Core-periphery structure in networks (revisited). SIAM Rev. 59(3), 619–646 (2017)
10. Holland, P. W., Laskey, K. B., Leinhardt, S.: Stochastic blockmodels: First steps. Soc. Netw. 5(2), 109–137 (1983)

Wire together, survive together: Structural stability and signal for collapse of interaction networks.

Sooyeon Yoon[1], Alexander V. Goltsev[1,2], and José F. F. Mendes[1]

[1] Department of Physics & I3N, University of Aveiro, 3810-193 Aveiro, Portugal
syoon@ua.pt
[2] A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

1 Introduction

The impact of negative external factors such as catastrophic environmental changes on biological and social networks can lead to a collapse of the systems when interactions between subjects forming the systems cannot resist anymore the factors [1]. It is well recognized that the structure plays a very important role in the robustness of complex systems against random and targeted attacks [2, refs. therein]. The big questions in complex systems science are what causes a collapse of some systems, how to predict the approach to the tipping point, what is the role of network structure in the stability of real complex systems [3]. One important characteristics of network structure is the network cohesiveness. The 'k-core', introduced to characterize the cohesion in social networks [4], is the largest subgraph whose all vertices have, at least, k nearest neighbors. Since the k-cores represent the most connected part of a network, one would expect that the k-core organization might play an important role in the structural stability of real complex network against damages and negative external factors. In this study, we explore the role of topology and the heterogeneity of interactions in the structural stability of networks of positively interacting agents subjected to a negative external field, which suppresses the activity of the agents. We study how positively interacting agents support each other to confront the negative field and the role of the k-core organization in the structural stability of the interacting system. In our approach, we understand the structural stability as the existence of a giant connected component of the network of active agents stable against perturbation. We also develop a new method of structural analysis based on a statistical analysis of so-called 'corona' clusters belonging to k-cores. This method allows us to reveal structural changes in the k-core organization when increasing the negative external field and allow to predict the collapse of weighted and unweighted networks. Structural stability of some real networks against negative external fields is also discussed.

2 Results

We introduce the energy of the system which governs the agents states ($x_i = 1$, if i is active and $x_i = 0$, if i is inactive.) as follows.

$$E = -\frac{1}{2}\sum_{ij} w_{ij} A_{ij} x_i x_j - \sum_{i=1}^{N} U_i x_i \,, \tag{1}$$

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

where w_{ij} is a weight of the link between i and j, A_{ij} is a component if adjacency matrix, and U_i is a strength of the external fields. We show that critical changes in the structure of interaction networks precede the network collapse. Nodes of degree q equals to the k-core index (i.e., $q = k$) at $k \geq 3$ play a special role in structural stability of the k-core. These 'corona' nodes form 'corona' clusters inside the k-core. If a 'corona' node belonging to a 'corona' cluster is removed then all other nodes belonging to the same 'corona' clusters are also removed one by one (the domino effect) because their degrees become less than k. It is the mechanism of avalanches that destroys the k-core at the tipping point [5]. We introduce a parameter,

$$\chi_{cr}(k) = \frac{\sum_\alpha s_\alpha^2(k)}{\sum_\alpha s_\alpha(k)} = \sum_\alpha \pi_\alpha s_\alpha(k), \tag{2}$$

where $s_\alpha(k)$ is the size of a 'corona' cluster with index α in the k-core. $\pi_\alpha \equiv s_\alpha(k)/\sum_\alpha s_\alpha(k)$ is the probability that a randomly chosen corona node in the k-core belongs to a corona cluster α. The parameter $\chi_{cr}(k)$ has a meaning of the mean size of corona clusters to which a randomly chosen corona nodes belongs. At the critical point of k-core collapse the parameter $\chi_{cr}(k)$ diverges in the limit $N \to \infty$. Thus, the tipping point of the k-core collapse is the percolation point of the 'corona' clusters. Based on these results we propose the following method, which allows to reveal structural changes of the interaction network that occur when approaching the tipping point. For each value of a control parameter, which can be either the field strength, the fraction of removed agents, time, or temperature, we find k-cores by use of the pruning algorithm and statistics of corresponding corona clusters. If $\chi_{cr}(k)$ increases when increasing (or decreasing) the control parameter then it means that the system approaches a point at which the k-core collapses. Our analysis of the k-core organization reveals that the observed dependence of the fraction M of the active agents versus $|U|$ is due to a non-monotonous decrease of sizes of k-shells for $2 \leq k \leq k_h - 1$. The k-shells decrease in size but do not disappear completely in contrast to the sequential collapse of k-shells in unweighted networks. At field strength above the critical point of the collapse of the k_h-core, the sharp decrease of M is caused by sequential breakdown of the remaining k-cores, starting at $(k_h - 1)$-core and finishing at 2-core. The results of our simulations demonstrate that unweighted and weighted networks are destroyed in opposite way by the negative field. In unweighted networks, the negative field first destroys the lowest 2-core and then one by one all higher cores, in contrast to weighted networks where the process of the destruction goes in the opposite order: first the highest k-core is destroyed, and then the field destroys one by one the lower cores up to 2-core. The origin of this difference is due to the difference in the pruning processes. In unweighted networks, we prune only nodes with degree not larger than a given threshold. In weighted random networks, a removed node can have an arbitrary degree. There is only one restriction: the total weight of all edges must be smaller than the threshold. The fact, that degree of the removed node is arbitrary, makes this pruning process to be similar to the removal of nodes at random independently on their degree. We applied our model to two real networks in ecosystems to analyze the structural stability of this kind of network against external negative factors: the unweighted network of plants and pollinators, and the weighted network

of below-ground plants-fungus symbioses (plant roots are colonized by diverse fungi, which increase host plant fitness).

Summary. We studied the structural stability of weighted and unweighted networks of positively interacting agents against a negative external field. We showed that positively interacting agents support the activity of each other and confront the negative field, which aims to suppress the activity of the agents. In our approach, we understand structural stability as the existence of a giant connected component of the network of active agents stable against perturbations. The competition between positive interaction and the negative field shapes the structure of stable states of the networks. In the case of unweighted (uniform interactions) networks, we demonstrate that the tipping point of network collapse caused by a strong negative field or weak interaction is determined by the highest k-core. With increasing the field strength or decreasing the interaction strength the network of active agents undergoes a cascade of transitions from k-core to $(k+1)-$core ground state. The field destroys from the small k-core and the highest k_h-core is destroyed at last. In weighted networks (heterogeneous interactions), the interplay between the topology and the distribution of weights determines the structural stability against a negative field. Our approach is based on pruning of agents which have insufficiently strong interaction (smaller than a threshold determined by the negative external field) with active agents. The advantage of this approach is that it takes into account not only direct, but also indirect impact of leaving agents on network structure. A leaving agent stops the interaction with other agents and weakens the strength, which holds these agents in the network. As a result, some of the agents can also be forced to leave the network. The cascade of these events can lead to a dramatic collapse of the entire network like a process when extinction events tend to trigger coextinction cascades of related species. We also demonstrated that a critical change in the structure of the system precedes the k-core collapse. This structural change is characterized by rapid growth of 'corona' clusters, signalling the approach to the tipping point. These structural changes create grounds for long-lasting avalanches and critical slowing down. They can serve as early warnings of the collapse.

References

1. Rothman, D. H.: Thresholds of catastrophe in the Earth system, Science Advances 3(9), e1700906 (2017)
2. Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F.: Critical phenomena in complex networks, Rev. Mod. Phys. 80, 1275 (2008)
3. Scheffer, M. *et al*: Anticipating critical transitions, Science 338(6105), 344–348 (2012)
4. Seidman, Stephen B: Network structure and minimum degree. Soc. Netw. 5, 269–287 (1983)
5. Baxter, G. J. and Dorogovtsev, S.N., Lee, K-E, and Mendes, J. F. F., and Goltsev, A. V.: Critical dynamics of the k-core pruning process, Phys. Rev. X 5(3), 031017 (2015)

The central role of peripheral nodes in directed network dynamics

Edgar A. P. Wright[1], Sooyeon Yoon[1], António L. Ferreira[1] José F. F. Mendes[1], and Alexander V. Goltsev[1,2]

[1] Departamento de Física & I3N, Universidade de Aveiro, Aveiro, Portugal,
wrighteap@ua.pt
[2] A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

1 Introduction

Many social, technological, and biological systems with asymmetric interactions display a variety of collective phenomena, such as opinion formation and synchronization. This has motivated much research on the dynamical impact of local and mesoscopic structure in directed networks. Here, we control the global organization of directed Erdős–Rényi networks, and study its impact on the emergence of synchronization and ferromagnetic ordering, using Kuramoto and Ising dynamics. In doing so, we demonstrate that source nodes – peripheral nodes without incoming links – can disrupt or entirely suppress the emergence of synchronized and magnetized states in directed networks [1]. In general, source nodes are a structural feature of directed networks, and may therefore have a significant impact on other dynamics with local pairwise interactions, given a specific set of dynamical parameters, link properties, and initial conditions, as well as the local and mesoscopic structure of the network.

2 Structure

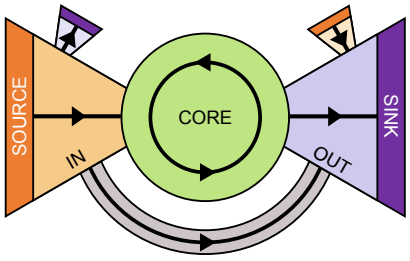


Fig. 1. Bow-tie (core-periphery) architecture of directed networks. The CORE is the largest strongly connected component (nodes reachable from each other through a sequence of directed links). The periphery comprises the IN and OUT components – nodes in sequences of directed links leading into and out of the CORE respectively – and a hierarchy of tendrils [2]. As indicated by the arrowheads, the overall connectivity of this architecture is feedforward, from SOURCE nodes (IN nodes without in-links) to SINK nodes (OUT nodes without out-links).

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

A general strategy for characterizing the global organization of directed networks was first applied to the World Wide Web [3], revealing the bow-tie architecture schematically depicted in Fig. 1. Here, we consider a toy-model of bow-tie organization based on directed Erdős–Rényi networks with $N = 10^5$, where each node has an incoming link with probability $2\langle q_{in}\rangle/(N-1)$, and the relative number of nodes and links in each bow-tie component is determined by the mean in-degree $\langle q_{in}\rangle$ [4, 1].

3 Dynamics

We investigate how the emergence of synchronization and magnetization in the Kuramoto model (KM) and the Ising model (IM) are impacted by structure, controlling both the structural parameter – the mean in-degree $\langle q_{in}\rangle$ – and the dynamical parameters – the coupling strength K (KM) and the temperature T (IM) – see [1] and references therein for dynamical equations and methods. In the KM, the macroscopic state of N oscillators with phase θ_n is described by the order parameter (complex amplitude)

$$r = \frac{1}{N}\sum_{n=1}^{N}\cos\left(\theta_n - \psi\right), \tag{1}$$

where $r = 1$ corresponds to full synchronization, and the average phase $\psi = \sum_n \sin(\theta_n)/\sum_n \cos(\theta_n)$. In the IM, the macroscopic state of N spins s_n is described by the order parameter (magnetization per spin site)

$$m = \frac{1}{N}\sum_{n=1}^{N} s_n, \tag{2}$$

where $m = 1$ corresponds to full magnetization. The response of the system to structural changes is then captured by the corresponding pair correlation functions

$$C = N\left[\langle r^2\rangle_t - \langle r\rangle_t^2\right], \tag{3}$$

and

$$\chi = N\left[\langle m^2\rangle_t - \langle m\rangle_t^2\right], \tag{4}$$

where $\langle\ \rangle_t$ denotes a time-average.

4 Results

In the limit where $K \gg 1$ and $T \to 0$, we find that the macroscopic state of the system is determined by the network's bow-tie architecture. By studying the system's response to the removal of a fraction of IN nodes f_{IN}, selected uniformly at random, we further find that the internal dynamics of SOURCES, which drive IN dynamics, act to disrupt the emergence of synchronization and ferromagnetic ordering in the CORE, as shown in Fig. 2.

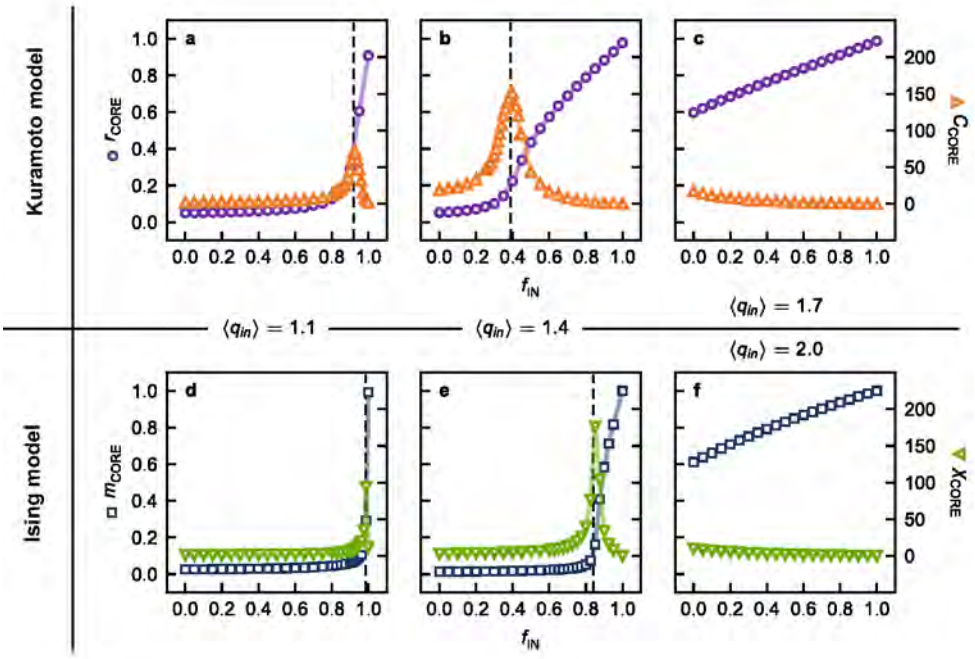


Fig. 2. Synchronization r_{CORE} (circles) and magnetization m_{CORE} (squares) in the Kuramoto – (a), (b), (c) – and Ising – (d), (e), (f) – models, and the corresponding pair correlation functions C_{CORE} (up triangles) and χ_{CORE} (down triangles) in the CORE of directed Erdős–Rényi networks with mean in-degree $\langle q_{in} \rangle$, as a function of the fraction of randomly removed IN nodes f_{IN}. The values of $\langle q_{in} \rangle$ are 1.1 in (a) and (d), 1.4 in (b) and (e), 1.7 in (c), and 2.0 in (f). Dashed vertical lines indicate peaks in C_{CORE} or χ_{CORE}. For details on the calculation methods see [1].

References

1. Edgar A. P. Wright, Sooyeon Yoon, António L. Ferreira, José F. F. Mendes, and Alexander V. Goltsev. The central role of peripheral nodes in directed network dynamics. *Scientific Reports*, 9(1):1–11, September 2019.
2. G. Timár, A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes. Mapping the Structure of Directed Networks: Beyond the Bow-Tie Diagram. *Physical Review Letters*, 118(7):078301, 2017.
3. Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. *Computer Networks*, 33(1):309–320, 2000.
4. S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Giant strongly connected component of directed networks. *Physical Review E*, 64(2):025101, 2001.

Effective Connectivity vs. Average Sensitivity: Representing Polyadic Relationships in Models of Complex Networks

Manuel Marques-Pita[1,3] and Luís M. Rocha[2,3],

[1] Cicant, ULHT, 388 Campo Grande. Lisbon, Portugal
[2] Center for Social and Biomedical Complexity, School of Informatics, Computing and Engineering. Indiana University, Bloomington IN, USA.
[3] Instituto Gulbenkian de Ciência. Oeiras, Portugal
manuel.pita@ulusofona.pt, rocha@indiana.edu

Introduction

Network models have become an essential tool to study biological, social and technological complex systems. The study of structural network properties has yielded deep insights about their organization and function [3]. Yet, there is growing interest in studying *structure* and *dynamics* together [see e.g. 1, 11, 14], particularly, in the context of studying control in complex networks—which we know does not depend on structure alone [7]. Boolean Networks (BNs) are canonical models of complex systems in which node behaviour can be easily described using ON-OFF automata. BNs have been used successfully to model a wide range of real-world systems [e.g. 8, 11, 17]. They have also been used to study the structural and node-behaviour parameters that determine network *dynamical regimes* [9].

A complex network can be in one of three dynamical regimes: *stable, unstable* or at the *critical* edge between the two. Living, social and technological systems are believed to be in this latter, critical, regime [5, 9]. Derrida and Pomeau [6] were the first to attempt relating network structure and node behaviour to the dynamical regime of BNs. Their theory predicts the dynamical regime of BNs, based on fixing the BN in-degree and the bias of the automata functions in their nodes. It defines the boundary between stable and unstable (chaotic) regimes as $2kp(1-p)=1$ [see 6]. The bias is a measure of how skewed an automaton is to one of its output states—computed as the probability that the automaton transitions to ON. Even when the fixed in-degree assumption is relaxed, and as long as the in-degree distribution has a characteristic mean value $\langle k \rangle$, this theory has the same predictive value [2]. However, this theory often mispredicts the dynamical regime of BNs near the critical edge [10].

One of the key mechanisms that plays a role in critical network dynamics is *canalization,* which was defined to explain how phenotype traits are conserved when there is vast genetic variation in living systems [16]. Canalization can be easily studied using automata networks [9]. Consider for instance an automaton f that transitions to ON using the logical OR function of its x inputs. It is clear that f will always transition to ON, as long as at least any of its inputs is ON, and *regardless* of the state of the other $x-1$ inputs (see Figure 1A). Canalization in automata can either be a (strict) dyadic

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

input to output relationship, where the state of one input of f alone is always sufficient to determine its transition, or a (collective) polyadic relationship where the states of two or more inputs are needed, together, to determine the automaton's transition [12]. In our previous example, the LUT entry where all inputs are zero is the only polyadic (collective) relationship in f (the OR logical function).

We introduced a measure of canalization in automata, *effective connectivity* (k_e^f) that tallies the expected number of input states that are *sufficient* to determine the transition of f [11]. The effective connectivity of an automaton f is computed from its set of wildcard schemata (see Figure 1A and [11] for details). A related measure, *average sensitivity*, [13] was used recently as predictor of the dynamical regime of a large set of Boolean network models of biochemical regulation and signalling [5]. To obtain the *average sensitivity,* (s^f), of an automaton f, we must first get the *partial derivatives* of f with respect to each input x_i. Then we can compute the *activity* of every input x_i in f, which is the probability that changing its state will change the automaton's transition—computed from the partial derivatives. Finally, s^f, is the average of the input-variable activities . The average sensitivity was defined to be a predictor of dynamic regime in BNs [13], and the transition between stable and unstable regimes has been observed when $s^f = 1$ [see 15].

Here we compare *average sensitivity* [13] and *effective connectivity* [11], showing that the latter is a more complete representation of canalization in the effective logic of automata. We also show that effective connectivity is likely to lead to better predictions of the dynamical regime of BNs than those obtained from average sensitivity.

Results

A key difference between s^f and k_e^f is that the schemata from which k_e^f is computed, allow the representation of the original BN as a bipartite network, threshold network or hypergraph. Such representations not only capture the effective structure of the constituent automata nodes in BNs, but this structure is well described by the corresponding k_e^f values [11]. Average sensitivity s^f cannot be readily translated to a similar network structural parameter or related graph representation. It is computed directly from the automaton's LUT by averaging over the accumulated the effects of perturbations on single inputs. This means that s^f does not explicitly differentiate the effect of a single input perturbation on a dyadic or polyadic input-output relationship in the automaton.

Concerning the predictability of dynamic regime, we produced an ensemble $\mathscr{S}$ containing 1000 unique automata with $k = 7$ inputs and bias $p = 0.07$. Having the same bias means that all automata in $\mathscr{S}$ have the same number of 1s in their LUT output list. Dyadic input-output relationships are often seen when groups of input-state combinations, which are next to each other in the automaton's LUT, have the same state transition (output). This assumes input-state combinations are ordered lexicographically in the LUT (like in Figure 1A). The opposite is also generally true: if e.g. the 1s in the output list of an automaton f LUT are scattered in the LUT, there is a higher probability that f has more polyadic input-output relationships (collective canalization), see Figure 1B. Since, according to the current theory [6], the critical bias for $k = 7$ is $p_c = 0.077$, using these parameters should in principle produce networks that at the critical edge. Figure

1B depicts the result of computing the average sensitivity and effective connectivity for the automata in $\mathscr{S}$. For effective connectivity, median$(k_e^f) = 2.08$; IQR$(k_e^f) = 0.65$, while for average sensitivity, median$(s^f) = 1.07$; IQR$(s^f) = 0.22$. Note that while most automata in $\mathscr{S}$ are sensitive to around one input, effective connectivity shows various regimes where some automata depend on average on 1.3 inputs, and others on more than 3 inputs. This is because (k_e^f) accounts for polyadic relationships (collective canalization) in automata, which is ultimately a more realistic characterization of their true behaviour. The prediction of dynamical regime from average sensitivity would be that most BNs built from automata in $\mathscr{S}$ will be unstable, since their average sensitivity is $s^f \geq 1$ [13, 15]. In contrast, we show that BNs built from LUTs in the yellow area of Figure 1B exhibit different dynamics from BNs built from LUTs in the pink area (see Figure 1C). These results show that effective connectivity better characterizes collective canalization (polyadic relationships) and thus is likely a better predictor of dynamical regime. Many of the networks predicted by the average sensitivity to be unstable are in the critical and stable regimes.

Finally, as [4] points out, many real world networks are made of polyadic rather than purely dyadic relationships. These are better modelled with hypergraphs which are a natural representations of effective connectivity and wildcard schemata [11]. Novel methods to apply classical network science insights and tools are being currently studied for hypergraphs, along with methods to generate random hypergraphs with desired properties [4]. This will allow us to study representations that capture both structure and dynamics in network models considering that local interactions can be, and often are, polyadic in the real world.

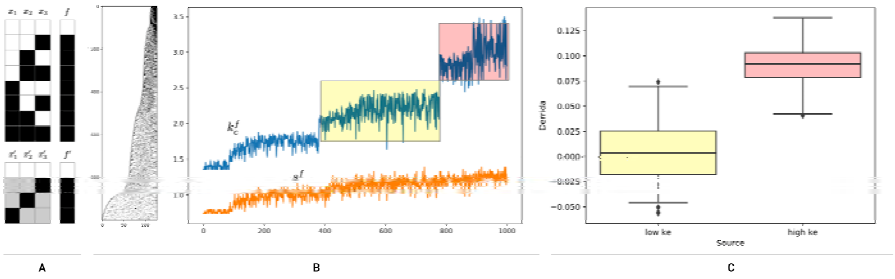


Fig. 1. (A) Schemata for the LUT of the 3-input OR function, $k_e^f = 1.25$. (B) 1000 $k = 7$ LUTs with bias $p = 0.07$ were chosen by increasingly scattering the 1s in the output column. Next to this is the plot showing the corresponding k_e^f and s^f. Notice that for very similar values of s^f the corresponding k_e^f can change significantly (yellow and pink shaded areas). (C) Finally, the Derrida coefficients of fixed in-degree $k = 7$ BNs with $N = 32$ nodes, built using LUTs in the yellow shaded area of (B) are in the critical regime, while BNs built from LUTs int the pink shaded area of (B) are slightly into the chaotic regime.

Bibliography

[1] R. Albert. Boolean modeling of genetic regulatory networks. In *Complex networks*, pages 459–481. Springer, 2004.

[2] M. Aldana. Boolean dynamics of networks with scale-free topology. *Physica D: Nonlinear Phenomena*, 185(1):45–66, 2003.

[3] A.-L. Barabási. The network takeover. *Nature Physics*, 8(1):14, 2011.

[4] P. S. Chodrow. Configuration models of random hypergraphs and their applications. *arXiv preprint arXiv:1902.09302*, 2019.

[5] B. C. Daniels, H. Kim, D. Moore, S. Zhou, H. Smith, B. Karas, S. A. Kauffman, and S. I. Walker. Logic and connectivity jointly determine criticality in biological gene regulatory networks. *arXiv preprint arXiv:1805.01447*, 2018.

[6] B. Derrida and Y. Pomeau. Random networks of automata: a simple annealed approximation. *EPL (Europhysics Letters)*, 1(2):45, 1986.

[7] A. J. Gates and L. M. Rocha. Control of complex networks requires both structure and dynamics. *Scientific reports*, 6, 2016.

[8] T. Helikar, B. Kowal, S. McClenathan, M. Bruckner, T. Rowley, A. Madrahimov, B. Wicks, M. Shrestha, K. Limbu, and J. A. Rogers. The cell collective: toward an open and collaborative approach to systems biology. *BMC systems biology*, 6(1):96, 2012.

[9] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Genetic networks with canalyzing boolean rules are always stable. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17102–17107, 2004.

[10] S. Mannika, M. Marques-Pita, and L. M. Rocha. The effective connectivity of biochemical networks determines their critical dynamics. *(submitted to PNAS)*, 2019.

[11] M. Marques-Pita and L. M. Rocha. Canalization and control in automata networks: body segmentation in drosophila melanogaster. *PloS one*, 8(3):e55946, 2013.

[12] C. O. Reichhardt and K. E. Bassler. Canalization and symmetry in boolean models for genetic regulatory networks. *Journal of Physics A: Mathematical and Theoretical*, 40(16):4339, 2007.

[13] I. Shmulevich and S. A. Kauffman. Activities and sensitivities in boolean network models. *Physical review letters*, 93(4):048701, 2004.

[14] S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2018.

[15] M. Villani, D. Campioli, C. Damiani, A. Roli, A. Filisetti, and R. Serra. Dynamical regimes in non-ergodic random boolean networks. *Natural Computing*, 16(2):353–363, 2017.

[16] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, 150(3811):563, 1942.

[17] R.-S. Wang and R. Albert. Elementary signaling modes predict the essentiality of signal transduction network components. *BMC systems biology*, 5(1):1, 2011.

Part XX

Urban Networks

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Graph-based Inference from Non-Probability Road Sensor Data

Jonas Klingwort[1,2], Bart Buelens[3], Joep Burger[1], and Rainer Schnell[2]

[1] Statistics Netherlands, CBS Weg 15, 6412 EX Heerlen, The Netherlands
j.klingwort@cbs.nl, jonas.klingwort@uni-due.de
[2] University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany
[3] Flemish Institute for Technological Research, Boeretang 200, 2400 Mol, Belgium

1 Introduction

Big data has received increasing attention across several disciplines in recent years. In Official Statistics, big data potentially enables us to produce statistics cheaper, faster, and on a higher level of detail [1]. In contrast to traditional probability samples, however, big data typically lacks a sampling design. The unknown data generating mechanism makes design-based inference methods inapplicable. Therefore, research on methods to use big data for population inference in official statistics is required. In this paper, we aim to infer the truck traffic distribution in the Dutch road network from sensors installed on a non-probability sample of road segments. Our study is an illustration of making inference from data without a sampling design.

2 Data and Methods

Data We use real-world data from a sensor system installed at 18 sensor stations on Dutch national freeways to detect overloading. When a truck passes a sensor station, it is weighed, classified, and a photograph of the front license plate is taken. Using the license plate as a unique identifier, we could link information about the truck and owner from the vehicle and enterprise registers. Out of the 36 million trucks recorded in 2015, 15 million could be linked. Since trucks could pass a station multiple times a day, the number is reduced to 14 million unique trucks. In this proof of concept, we restrict the data to a single week with the largest number of simultaneously working sensors. Other weeks will be added later to increase mass and to borrow strength from adjacent days.

In the second step, the Dutch transport network of freeways was constructed by scraping interchange road junctions (the vertices) and their connecting freeways (the directed edges) from www.wegenwiki.nl. Six vertex features were computed: degree, strength, betweenness, closeness, vulnerability and clustering coefficient, using the inverse haversine distance between vertices as the edge weight. For realistic vertex feature values, the network was expanded with neighboring freeways in Belgium and the German states North Rhine-Westphalia, Lower Saxony and Bremen. Figure 1 shows the resulting graph. The Dutch part consists of 108 vertices and 284 edges. The 18 measuring stations of the sensor network were assigned at 18 edges of the graph using their geolocation.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

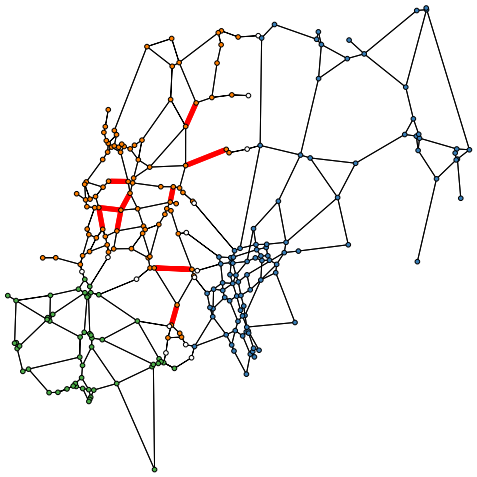


Fig. 1. The transport network in geographical layout: Dutch (orange), Belgian (green), and North-West German (blue) interchanges, connected by freeways. White vertices are border crossings. Red edges are highways with installed measuring stations.

Methods The probability of a truck driving on an edge of the Dutch freeway network on a given day is modeled using a GLM with logit link and binomial error distribution. The relationship between on the one hand the relative number of trucks detected by a sensor installed on one of the 18 edges is modeled, and on the other hand the features of the origin vertex and the destination vertex. The used features are: weight (inverse edge length (km^{-1}), degree, strength, betweenness, closeness, vulnerability, and a cluster coefficient. The relationship is used to predict the detection probability for all trucks in the vehicle register for all network edges. The modeled probabilities are multiplied with the number of trucks registered in the vehicle register constituting the study population to derive the edge counts.

3 Preliminary results

Figure 2 shows the predicted distribution of the number of trucks driving on the Dutch road network on a given day. The graph, weighted by the modeled edge counts, allows identifying routes to neighboring countries (Germany, Belgium), while truck traffic in parts of the center and the western country seems to be lower. This was expected as the studied population includes a large proportion of heavy traffic trucks. Such vehicles often take routes from the seaports (west Netherlands) through the Netherlands and further into Europe. The relationship between predicted and actual edge counts is according to Pearson correlation coefficient very strong ($r = 0.97$). Moreover, a low RMSE of 141 is achieved. The cross-validation results show, however, a weak linear relationship between the predicted and true actual edge counts ($r = 0.35$) and the RMSE increases to 2581. Hence, the predictive model fits well but does not yet generalize well yet. Potential improvements will be discussed in the next section.

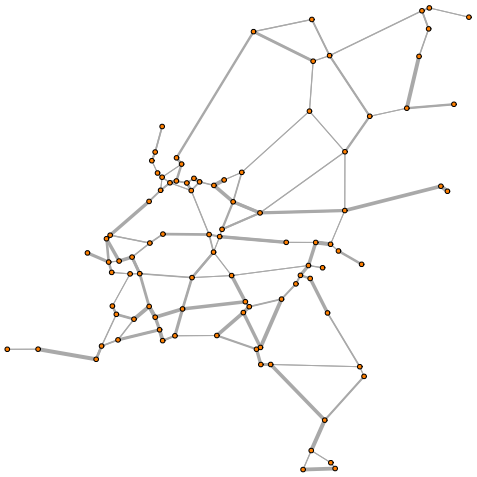


Fig. 2. Estimated traffic distribution of the number of trucks in the Dutch road network. The thickness of edges corresponds to the estimated number of trucks passing.

4 Conclusions and Future research

We demonstrated a method to use big data in official statistics. It can be used to infer the traffic distribution in a transport network from sensors installed on a non-probability sample of edges. Such a method requires a unique identifier to identify elements in a big data source belonging to the inferential population. This method is based on the assumption that data is missing at random.

The proposed methodology can be expanded in several ways. First, the edge feature set will be extended with traffic intensity data from a more extensive road sensor system, consisting of 24 thousand sensors, but without cameras to identify trucks [2]. Data exploration showed similar time series of traffic patterns, which should make these data a good predictor to model the edge counts. Second, we will use several register features about the truck and owner to stratify the analysis. Third, the dataset will be extended with the entire time series. Time series modeling will be applied to account for the dependency between days. Both stratification and adding days will reduce the risk of overfitting. Finally, we consider using an open Jackson network to account for the spatial dependency between edges.

References

1. Piet J. H. Daas, Marco J. Puts, Bart Buelens & Paul A. M. van den Hurk. 2015. Big Data as a Source for Official Statistics. Journal of Official Statistics. 31 (2), 249–262
2. National Data Warehouse for Traffic Information. 2016. NDW: A Nationwide Portal for Traffic Information. Retrieved from: `https://www.ndw.nu/documenten/nl/`

Stars of mitigation? Participation-based structure in a city-to-business network

Milja Heikkinen[1,2], Onerva Korhonen[3], Sirkku Juhola[1,4], and Tuomas Ylä-Anttila[5]

[1] Ecosystems and Environment Research Programme, University of Helsinki, Finland
[2] Helsinki Institute of Sustainability Science (HELSUS), Finland
[3] Université de Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, Lille, France,
onerva.korhonen@gmail.com,
WWW home page: onervakorhonen.wordpress.com
[4] Centre for Climate Science and Policy Research, Linkping University, Sweden
[5] Faculty of Social Science, University of Helsinki, Finland

Climate change is one of the most important problems the humankind is facing. Cities are considered to play a key role in climate change mitigation [1]. However, private actors are responsible for creating most of the greenhouse gas emissions, and reducing these emissions is a central target of climate governance [2]. Therefore, when public authorities, for instance cities, want to reduce emissions, they need to cooperate with the private sector.

Here, we address the case of the Climate Partners (CP), a city-to-business network initiated by the city of Helsinki, Finland. Founded in 2011, CP connects 83 companies located at the Helsinki metropolitan area. CP maintains two lines of action. First, companies joining the network sign a Climate Commitment that defines their individual goals related to climate change mitigation. Second, CP organizes seminars and workshops for the member companies. The aims of CP include introducing new operating methods and business opportunities, reducing emissions through cooperation, and sharing knowledge and experiences on best practices [3].

All the above-mentioned aims relate to collaboration between companies. To reach these aims, CP needs both to bring together companies from different fields of business and to get these companies engaged in the CP activities. To investigate how well CP meets these two requirements, we construct a bipartite network, where the bottom and top nodes are, respectively, the CP events organized from 2011 to 2018 and the companies that participated in these events. We recognize that companies can take actions to mitigate climate change also outside of the CP activities. However, such indpendent actions of companies are outside of the scope of the present work, as we specifically address the role of the city-to-business network in climate change mitigation.

To address the diversity of participating companies, we use the method suggested by by Makino & Uno [4] to detect the bicliques of the bipartite network. First, we obtain a monopartite network by adding a link between each pair of top and bottom nodes; cliques of this monopartite network include every biclique of the original network and two additional cliques corresponding the sets of top and bottom nodes. Then, we detect the cliques of the monopartite network by the NetworkX *find_cliques* function [5–7].

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

Each of the detected bicliques corresponds to a set of events and companies that participated in all these events. To quantify the diversity of the companies participating in same events, we first divide the companies to 20 fields of business. Then, we apply a diversity measure adopted from ecology, *effective diversity* D_{eff} [8, 9]. For clique A with N_f fields,

$$D_{eff} = \frac{1}{1 - GS(A)} = \frac{1}{\sum_{i=1}^{N_f} p_i^2}, \tag{1}$$

where $GS(A)$ is the Gini-Simpson index of clique A, or the probability that two companies randomly picked from A are from different fields, and p_i is the fraction of companies from field i out of all companies in A. In practice, $D_{eff}(A)$ tells the number of different fields in a population with the same Gini-Simpson index as A and with the fields equally distributed among participating companies. To compare the diversity between cliques of different sizes, we normalize D_{eff} by clique size to obtain *normalized diversity* D_n.

To address the engagement of companies in CP activities, we first define a *bi-star* as a (N_i,1) clique that contains an event and the N_i companies that participated in only this event. Then, we define the *starness* of the bipartite network G as

$$S(G) = \frac{\sum_{i=1}^{N_{stars}} N_i}{N_C}, \tag{2}$$

where N_C is the total number of companies.

The cliques of the CP network have companies from, on average, 6.04 different fields, leading to mean effective diversity of 5.39 and mean relative diversity of 0.91. The starness of the CP network is 0.57. To interpret these values, we define two null models: field-shuffled networks that retain the original link structure while the fields of companies are randomly re-distributed and link-shuffled networks where the companies and events are randomly re-connected. Comparison to the null models (Fig. 1) reveals that the diversity of the CP network is similar to that of the field-shuffled random networks (N_f: 6.04 vs 5.67, D_{eff}: 5.39 vs 4.80, D_n: 0.91 vs 0.88) , while the starness of the CP network is clearly higher than in the null model (0.57 vs 0.19).

In other words, CP manages to bring together companies from diverse fields, opening possibilities for information transfer and innovative collaborations. However, the engagement of companies in CP is low, which may make reaching CP's goals challenging. While the participating companies may see themselves as *stars of mitigation*, the CP bipartite networks merely consists of *bi-stars of mitigation*: companies that give up participating in CP activities after their first event.

The present work concentrates on the diversity and engagement of companies participating in CP events. In future, we will complement these results by analyzing the evolution of companies' Climate Commitments described in CP's annual reports. This analysis will help us to find out if CP membership has lead to more ambitious climate change mitigation actions even in companies that have decided to not continue participating in CP events.

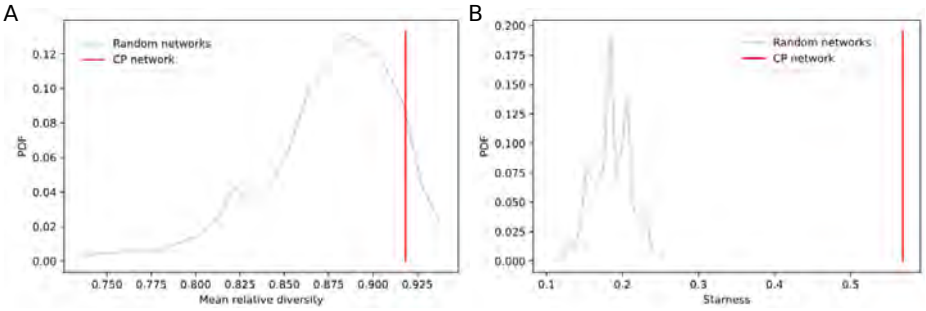


Fig. 1. Comparing the CP bipartite network to null models. (A) The mean effective diversity of the CP network is close to that of field-shuffled random networks. (B) The starness of the CP network is notably larger than that of link-shuffled random networks. Gray lines show the distributions of mean D_{eff} and S across 1000 random networks.

References

1. Revi, A., D.E. Satterthwaite, F. Aragn-Durand, J. Corfee-Morlot, R.B.R. Kiunsi, M. Pelling, D.C. Roberts, and W. Solecki (2014). Urban areas. In Field, C.B., Barros, V.R., Dokken, D.J., Mach, K.J., Mastrandrea, M.D., Bilir, T.E., Chatterjee, M., Ebi, K.L., Estrada, Y.O., Genova, R.C., Girma, B., Kissel, E.S., Levy, A.N., MacCracken, S., Mastrandrea, P.R., and White, L.L. (eds.) *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: Cambridge, UK and New York, NY, USA, 535–612.
2. Abbot, K. (2018). Orchestration. Strategic Ordering in Polycentric Governance in Jordan, A. (ed.) *Governing Climate Change. Polycentricity in Action*. Cambridge University Press: Cambridge, UK.
3. Climate Partners: https://www.ilmastokumppanit.fi/en/climate-partners/ (30.8.2019).
4. Makino, K. & Uno, T. (2004). New Algorithms for Enumerating All Maximal Cliques. In Hagerup T., Katajainen J. (eds) *Algorithm Theory - SWAT 2004*. SWAT 2004. Lecture Notes in Computer Science, vol 3111. Springer: Berlin, Heidelberg, Germany.
5. Bron, C. & Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* 16(9), 575–577.
6. Tomita, E., Tanaka, A., Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor. Comput. Sci.* 363(1), Computing and Combinatorics, 10th Annual International Conference on Computing and Combinatorics (COCOON 2004), 28–42.
7. Cazals, F. & Karande, C. (2008). A note on the problem of reporting maximal cliques. *Theor. Comput. Sci.* 407(1-3), 564–568.
8. Jost, L. (2006). Entropy and diversity. OIKOS 113(2), 363–375.
9. MacArthur, R. (1965). Patterns of species diversity. *Biol. Rev.* 40, 510–533.

Networks of Hospitals, Patients and Organ Donations within the US

Zachary Patterson, Sarah Richeson, and M. Abdullah Canbaz

Indiana University Kokomo,
2900 S. Washington St.,
Kokomo, IN, 46903
[zspatter,richessa,mcanbaz] @iu.edu

1 Introduction

Problems around the organ donation, utilization, transplantation, and operational infrastructure has been a discussion for decades in the United States. Between shortages of donor organs, complex structure of organ procurement agencies, hospital networks, and physicians, the organ donation problem has become a multi-player game where every party operates without being aware of the other parties' involvement and knowledge base [11]. Considering that the demand for the organs needed by patients increased by 15 fold in some cases [10], researchers and government agencies alike have become involved in this particular problem. In the United States alone as many as 20 people per day, both adults and children, die waiting on an organ because of issues with the transplant system. [9].

The potential increase in a recipient's chances for survival is highly related to the number of people registered as organ donors. In order to counter the problem, organizations spend most of their effort on increasing the list of potential donors. While organ donation is widely supported within the US with 95% of adults expressing support, a much smaller percentage, i.e. 58%, of adults are actually registered organ donors [10].

In addition, one deceased donor can potentially donate eight organs (i.e., a heart, liver, pancreas, intestines, two lungs, and two kidneys) [8, 3]. Living donors are able to donate one of their two kidneys, and also a portion of their liver [9]. Still, issues regarding compatibility, blood type matching, pediatric transplantation, and organ rejection ratios can make finding an appropriate donor difficult.

Furthermore, the maximum organ preservation period after harvesting the organ from a donor varies between 4 hours to 36 hours. For instance, hearts and lungs have the small maximum organ preservation time of between 4-6 hours. A livers maximum organ preservation time can be anywhere from 8-12 hours. The maximum organ preservation time for a pancreas can be anywhere from 12-18 hours. Kidneys have the longest maximum organ preservation time of 24-36 hours [5].

In order to deal with the current situation, the United States government supports 58 organ procurement organizations(OPOs) [4] which work to match donor organs to patients currently on an organ wait list. These organizations also provide a wide range of resources such as providing support for donor families [7]. While these organizations service different regions of the country, they coordinate primarily with neighboring organizations to utilize the organ preservation period efficiently.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

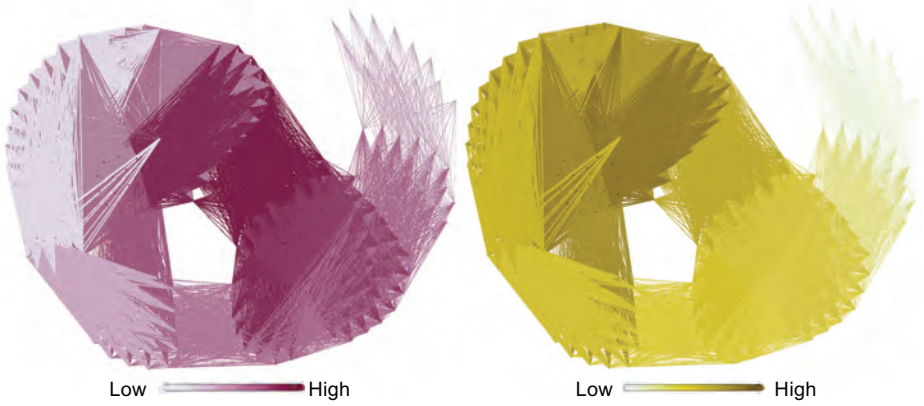


Fig. 1. Hospital Network clustered based on (left)Betweenness and (right) Eigenvector centrality

2 Outcomes

The United Network for Organ Sharing (UNOS), a U.S. Department of Health and Human Services initiative, maintains a centralized computer network which links all organ procurement organizations (OPOs), histo-compatibility labs, transplant hospitals, and transplant centers in a secure, real-time environment [1]. While this online system provides assistance in locating the donors for a particular type of organ needed in a certain case, it does not focus on optimizing the multi-player game between patient, organ, donor, and the physicians, such that even within a few block radius patients waiting for an organ might be missed because of the current system implementation [6].

In this paper, we analyzed the graph characteristics of current transplant hospital/center networks to better understand their implementation. To this end, we have collected wait lists, donation, and transplantation (including successful and failed transplants) data from UNOS. Based on the hospitals registered as the transplant locations within certain regions of the country, we have constructed the network of hospitals along with the registered patients and donors linked to each of these transplant facilities. While earlier studies have analyzed this problem from medical [13] and business [12, 14] aspects, to our knowledge, this is the first study to analyze the organ donation network.

UNOS [1] groups transplant facilities based on two criteria; the direct distance and regional weight. By utilizing the map of regions [10] and the distances of the cities [2], we have constructed the national organ transplant hospitals network of the United States.

Our first observation was that there is a rather densely connected graph respective to each sub-region of the graph and represents a given pair of nodes' favor-ability.

In Fig 1, we present the hospital network clustered by the betweenness centrality on the left and the eigenvector centrality on the right. We observe that the hospitals in east central U.S. are densely connected in both visuals. When considering the betweenness, we see that 47.35% of the hospitals, mostly the periphery states, are loosely connected

to network even though hospitals in these certain regions are connected to each other densely.

Looking at the **assortativity** metric, is a metric illustrating the preference of a node to link to others, We see high assortativity, 0.83. This is the artifact of having nodes connected to each other densely within their regions. Similar to this, we see the clustering coefficient to be positive, 0.34. Even though the graph assortativity is very high, the clustering coefficient reveals that there are not many cliques within the regions.

The most important aspect of this problem is that the actual network of hospitals that exist in the U.S. are the hospitals operated by different businesses in each state. Assortativity and the clustering coefficient clearly reveal that in the virtual level, the network of organ transplant facilities is primarily governed by business networks implemented by the economic motivations.

References

1. Access unos unet system: Unet organ transplant web platform, https://unos.org/technology/unet/
2. Distance between cities calculator, https://www.distance-cities.com/
3. Living donation, https://www.donatelife.net/types-of-donation/living-donation/
4. Optn about data, https://optn.transplant.hrsa.gov/data/about-data/
5. Organ procurement and transplantation network, https://optn.transplant.hrsa.gov/learn/about-transplantation/how-organ-allocation-works/
6. Unos data and transplant statistics: Organ donation data, https://unos.org/data/
7. Find your local organ procurement organization (Apr 2019), https://www.organdonor.gov/awareness/organizations/local-opo.html
8. Living-donor transplant (Aug 2019), https://www.mayoclinic.org/tests-procedures/living-donor-transplant/about/pac-20384787
9. Organ donation and transplantation statistics: Graph data (Mar 2019), https://www.organdonor.gov/statistics-stories/statistics/data.html
10. Organ donation statistics (Mar 2019), https://www.organdonor.gov/statistics-stories/statistics.html
11. Young op-ed: Reform organ donation to save lives (Jul 2019), https://www.young.senate.gov/newsroom/press-releases/young-op-ed-reform-organ-donation-to-save-lives
12. Gimbel, R.W., Strosberg, M.A., Lehrman, S.E.: Cultural analysis of an organ procurement organization. Progress in Transplantation 11(4), 249–254 (2001)
13. Satel, S.: The waiting game. American enterprise Institute for Public Policy Research, June 26 (2006)
14. Skaro, A.I., Hazen, G., Ladner, D., Kaplan, B.: Organ transplantation: an introduction to game theory. Transplantation 99(7), 1316–1320 (2015)

Mapping the Ecologies of the Dutch Energy Transition Hyperlink Network

Nuccio Ludovico[1], Franco Ruzzenenti[2], and Marc Esteve Del Valle[3]

[1] Sapienza Universit di Roma, Department of Psychology of Developmental and Socialization Processes, Piazzale Aldo Moro, 5, 00185 Roma, Italy, and University of Groningen, Center for Energy and Enviornmental Sciences, Nijenborgh 6, 9747 AG Groningen, The Netherlands
nuccio.ludovico@uniroma1.it,
[2] University of Groingen, Center for Energy and Environmental Sciences, Nijenborgh 6, 9747 AG Groningen, The Netherlands,
[3] University of Gronignen, Center for Media and Journalism Studies, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands

1 Introduction

Internet facilitates connections between a range of actors with a stake in the energy transition, including governments, environmental organizations, media outlets and corporations [1,2]. These connections tease a hyperlink network [3,4] affecting publics access to the information on energy transition issues. Despite its societal relevance, however, the characteristics of this network remain understudied. We present the results of a methodology we developed to study the Dutch energy transition hyperlink network. Our data shows the existence of a highly centralized network -with few authorities [5]- in which the debate about the energy transition revolves around a reduced number of topics.

2 Results

Figure 1 shows the methodology we developed to study the ecologies of the Dutch energy transition hyperlink network. To build the network we employed the Issue Crawler [6] and longitudinally collected data (February-May 2019) from the interactions (hyperlinks) between 9 websites representing key Dutch actors (public institutions, private companies, media, etc) with a stake in the energy transition debate (Phase 1). Then all the references to the energy transition debate were extracted from the nodes websites (N= 2,042) and (when possible) from its related Twitter accounts (Phase 2). Next, we collected data on the nodes location and aggregated all the nodes texts (website and tweets) into a single corpus (Phase 3). Last, we carried out our social network and topic modelling (Factor Analysis) analyses (Phase 4).

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

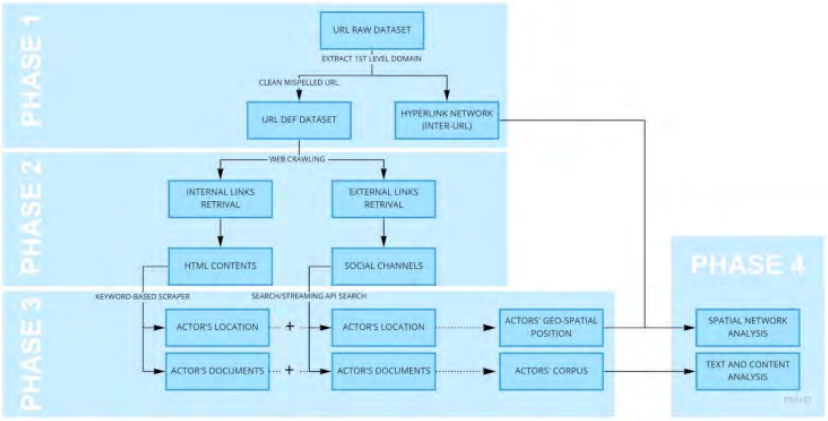


Fig. 1. A new methodology to study the Dutch energy transition hyperlink network.

Our results show the existence of a highly centralized network (with the nodes' degree fitting the power law distribution) in which few authorities concentrate most of the communication flows (see Figure 2). Of relevance, it is the leading role played in the network by some private companies (e.g. Siemens), public institutions (e.g.Netherlands Organization for Scientific Research -NWO-) and civil society organizations (e.g. Netherlands Wind Energy Association -NWEA-).

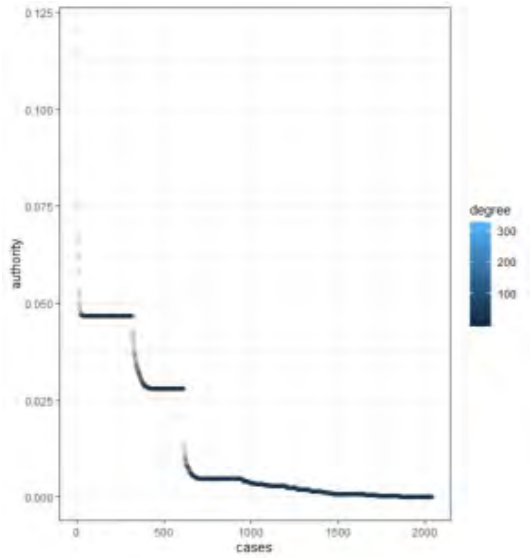


Fig. 2. Distribution of the communication flows in the Dutch energy transition hyperlink network. The chart shows that the degree distribution and the authorities distribution fit the power law.

Moreover, the results of our topic model reveal the existence of a reduced number of topics in the Dutch energy transition hyperlink network (see Figure 3).

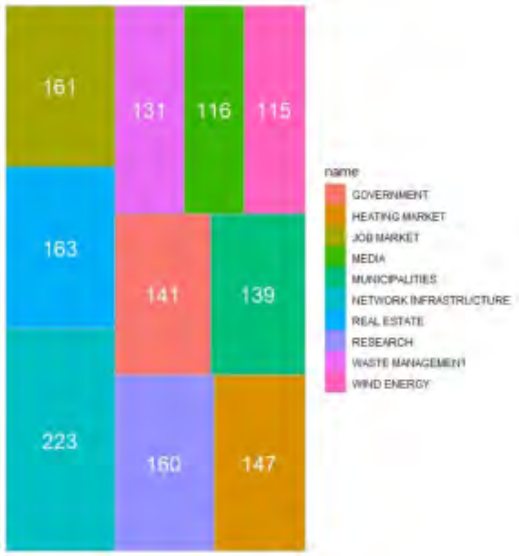


Fig. 3. Results of the topic model. The size of the squares indicates the number of terms allocated to each topic. Before cleaning the textual data, the corpus was of 116 documents and 11,727 unique words. Stop-words and terms with f¡ 5 were excluded of the analysis, resulting in a corpus of 116 documents and 746 unique words.

Particularly, Figure 3 shows the leading role of the ”Network Infrastructure” (N= 223) topic in the Dutch energy transition hyperlink network, followed by a set of topics related to the private sector, (”Real Estate”; ”Job Market”; and ”Heating Market”). A second group of topics are those related to the role of the Dutch national government and municipalities in the energy transition (”Government” and ”Municipalities”). Lastly, our model reveals the existence of three discussion topics linked to media outlets (”Media”), waste management (”Waste Management”) and wind energy (”Wind Eenergy”).

All in all, these findings reveal the existence of a Dutch energy transition hyperlink network in which few actors dominate the communication flows. Moreover, these communication flows revolve around a specific set of topics which seem to be led by market-oriented interests.

Summary. We have mapped the ecologies of the Dutch energy transition hyperlink network. Our results reveal the existence of a highly centralized network in which few authorities concentrate most of the communication flows. Indeed, the results of our topic model show the presence of a limited number of discussion topics.

References

1. Boykoff, M.: Who speaks for the climate? Making sense of media reporting on climate change. New York, NY: Cambridge University Press (2011)
2. Schäfer, M.S., Schlichting, I.: Media representations of climate change:A meta-analysis of the research field.Env.Com.8,142-160 (2014)
3. Park, H.W.: Hyperlink network analysis: A new method for the study of social structure on the web. Com. 25(1), 49-61 (2003)
4. Häussler, T.: Heating up the debate: Measuring fragmentation and polarisation in a German climate change hyperlink network. Soc.Net.54, 303-3013 (2018)
5. Kleinberg, J.M.: Authoritative sources in the hyperlinked environment. J. ACM. 46(5), 604-632 (1999)
6. Rogers, R.: Digital Methods. Massachusetts, MA: MIT Press (2013)

Does Road Network Topology Affect Real Estate Pricing? The Naples Case Study

Arianna Nocente[1], Jarir Salame Younis[1], Marco Cozzolino[1], and Giulio Rossetti[2]

[1] University of Pisa, Italy,
arcente@gmail.com, jarir.s.y@gmail.com,
marcocozzolino0@gmail.com
[2] KDD Lab, ISTI-CNR, Italy
giulio.rossetti@isti.cnr.it

1 Introduction

Nowadays, the estimation of house selling prices represent a hot and challenging task.The most widely used methods to address it rely on standard machine learning techniques, namely Artificial Neural Network and Hedonic. The former approach performs well with incomplete or unknown data [1], while the latter assumes that the selling price can be seen as a set of attributes and that the buyer tends to maximize its utility function [2, 3]. The main challenge of such a task lies in identifying those parameters that influence selling value since, often, they are interdependent or even hidden.

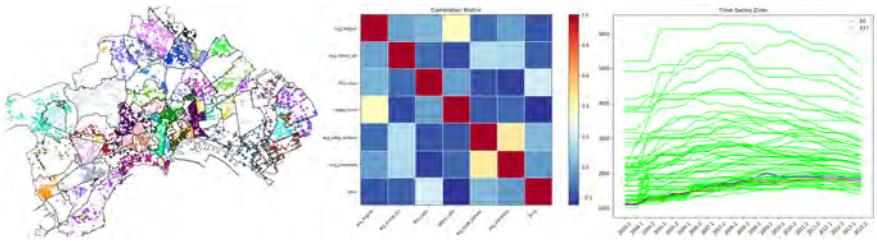
Using the road network of an Italian city, Naples (Figure 1(a)) – along with the dynamics of its points of interest –, our research aims to understand the relation between it and the real estate market prices. The road network was built identifying as edges the city roads and as nodes their intersections. Anomalies in the original data have been corrected through a cross-reference with the walk network. From a preliminary structural analysis emerges that the network shares some characteristics with subcritical random networks (e.g., the degree distribution) while at the same time, it does not exhibit the small-world property[4]. The degree distribution is close to the Gaussian distribution; planarity imposes severe constraints on the degree of a node and on its distribution, which is generally peaked around its average value[5].

We partitioned the network into sub-networks identifying the territorial boundaries of the 65 land registry areas[3]. For economic analysis, the network was enriched with main information about points of interests (POIs) which might influence the purchase choices of the housing stock made by population. The geospatial coordinates of such POIs have been identified and assigned to the nodes of the network which were nearest to them. A different approach has been used for educational and health institution POIs since the capacity and consequently the size of the universities and the hospital are not negligible. To take into account their relevance they have been mapped to areas and the nodes within them have been assigned to the same POI. The categories to which a POI belongs to and the number of nodes identified by them in the original network are, respectively: Public schools (70 nodes), Universities (21 nodes), Public and private

[3]The coordinates of these area boundaries were obtained through Geopoi, a cartographic visualization software for territorial navigation service.

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal



(a) Naples Road Network (b) Correlations (c) House Sales Estimates

Fig. 1: (a) Naples Road Network (with nodes subdivided by cadastral areas); (b) feature/pricing correlations; (c) pricing trends per area.

hospitals (186), Parks (34 nodes). The POI nodes are uniformly distributed among the identified subgraphs with the exception of the cadastral areas corresponding to industrial area and the city centre.

2 Preliminar Results

To investigate the correlations between estimates of sales prices, points of interest and structure of the network, each subgraph was characterized by the following network measures: (i) *Average degree*: average degree of nodes in the area; (ii) *Average street length*: the average value of the street length; (iii) *Average edge betweenness*: represent the importance of the streets and it is related to the number of commercial activities [6]; (iv) *Average closeness centrality*: related to the distance from any other node in the network; (v) *POIs Ratio*: number of POIs divided by the population of the area; (vi) *Edges Ratio*: number of streets divided by the population; this measure shows how well the street network serves citizens. The number of POIs was normalized using the expected population of the area to take into account the number of persons who would benefit from the services offered by them. The expected population of each cadastral area was estimated using the population and the borders of neighbourhoods assuming an equal distribution of the inhabitants inside neighbourhoods.

From our analysis, the Kendall and Pearson correlations among the identified features and the average cluster pricing appear close to zero (Figure 1(b)). Such a negative result seems to highlights how other features than the road network structural ones can explain the laws of the real estate market for Naples.

Indeed, one limit of the actual analysis lies in the absence of road network historical data. However, if such data were available, it would have been possible to train a predictive regressor that, taking into account how the city connectivity has evolved as well as when new POIs appeared, could have shown interesting predictive performances.

In absence of such information, we conducted a punctual survey by comparing the evolution of the price for the 65 Naples cadastral areas. To such extent, we leveraged real estate quotations data[4] over the decade from 2002 to 2013, detailed by semesters;

[4]Agenzia delle Entrate - Banca dati delle quotazioni immobiliare (OMI)

this data has been used to generate and compare the time series of each cadastral area (Figure 1(c)). For each area we computed the price linear trend; the trend coefficient shows how prices change over time. In particular, we focused our attention on the two areas where two important POIs (a hospital and a university) have been introduced during the observed period.

The average trend coefficient across all areas is 19.24. However, when considering the two selected areas the coefficient changes considerably reaching 35.84 in the new hospital area (Figure 1(c) red) and 41.42 in the new university one (Figure 1(c) blue). Such results underlying how for those areas the sales prices tend to grow twice as fast as in other areas of the city. Indeed, the evolution of the road network through the introduction of relevant POIs causes a strong evolution in prices. Moreover, such a result confirms what observed by the static analysis conducted so far: the prices in the cadastral areas where new POIs appear to change in an evident way, undergoing a rise, never become the highest prices in absolute terms, as they are the result of a historical evolution of more than ten years.

Dynamic analysis can be used as a tool to support forecasts combining the historical trend of prices with the evolution of the road network with a particular focus on the evolution of points of interest. As future work, we plan to collect data on the evolution of the Naples road network so to better formalize a dynamic graph theory framework able to explain the real estate market in terms of topology dynamics. In particular, integrating dynamic road network analysis with economic and cultural information (hospitals, schools, population) we are certain that a novel class of approaches for pricing forecasting can be devised. Moreover, such a framework could be used as a guide to identifying those areas for which urban re-qualification is needed or, in an entrepreneurial context, to choosing where to open a new POI to maximize the effects on sales prices.

Acknowledgment

This work is partially supported by the European Community's H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" grant agreement 654024, http://www.sobigdata.eu, "SoBigData".

References

1. Shaw, J.:Neural network resource guide. AI Expert. 8(2), (1992)
2. Griliches, Z.: Price Indexes and Quality Changes: Studies in New Methods of Measurement. Harvard University Press (1971)
3. Rosen S.: Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economics. 82, (1974)
4. Katzav, E. and Biham, O. and Hartmann, A. K.:Distribution of shortest path lengths in subcritical Erdős-Rényi networks. Phys. Rev. E. 98, (2018)
5. Barthlemy, M.: Spatial networks. Phys. Rep. 499(13), 1 – 101 (2011)
6. Porta, S. and Strano, E. and Iacoviello, V. and Messora, R. and Latora, V. and Cardillo, A. and Wang, F and Scellato, S.: Street centrality and densities of retail and services in Bologna, Italy. Environment and Planning B: Planning and design. 36(3). (2009)

On the Metropolis Algorithm for Urban Street Networks

Jérôme Benoit[1] and Saif Eddin Jabari[1,2]

[1] New York University Abu Dhabi, Saadiyat Island, POB 129188, Abu Dhabi, UAE
[2] New York University Tandon School of Engineering, Brooklyn, NY 11201, New York, USA
jerome.benoit@nyu.edu

1 Introduction

The complexity of urban street networks is well accepted to reside in the *information space* where roads map to nodes and junctions map to links [1–4]. By investigating the information space, we aim to provide new tools to study our cities.

Striking broad valence distributions have been observed among actual information networks. The urban street networks of self-organized cities deserve special attention for at least two reasons. First, they might have reached, over time, a spontaneous equilibrium that most designed cities fail to reproduce [5]. Second, the valence distributions of their information networks clearly distinguish themselves by following the (scale-free) Pareto distribution [2–4]. The second reason, while it supports the first one, may hopefully lead to a tractable theory.

Accordingly, we envision urban street networks as evolving social systems subject to a Boltzmann-mesoscopic entropy preservation [3, 4]. This preservation ensures the passage from the road-junction hierarchy to a scale-free coherence, i.e., that the valence distribution of the information network follows a Pareto distribution. The Boltzmann-mesoscopic entropies reflect the perception that inhabitants have of their own city, so they are better expressed in terms of *surprisal*. In brief, we conjecture that information networks tend to evolve by maintaining their amount of surprisal constant on average.

Even so information networks are well recognized as relevant, there is some art in how social and geographical criteria are pondered to construct them from urban street networks. However, the *deflection angles* between pairs of adjacent street-segments appear to be pertinent constructing parameters [2–4]. Naive *behavioural based join principles* [2] based on deflection angles have been used with good success [2]. Amazingly, the most successful one [2] is a random process with numerous outputs. The output arbitrariness must be addressed. Embracing the idea that information networks are driven by surprisal allows us to elevate these principles to a single-flip *Metropolis algorithm* as used for generating equilibrium states of Ising-like models in statistical physics [6].

2 Method

Here nodes are *natural roads* (or roads for short), that is, an exclusive sequence of successive street-segments joined according to some behavioural based join principle [2]. If beyond some threshold angle any joining has to be forbidden, multiple possibilities remain open. Two join principles based on deflection angle have appeared realistic

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

against well-founded cadasters. The *self-best-fit* and *self[-random]-fit* join principles operate sequentially on growing roads, until applicable, by randomly seeding them with a not-yet-selected street-segment before recursively appending, until applicable, one of the not-yet-appended street-segments. The self join principles differ only by the choice of the nominees. The self-best-fit join principle picks the not-yet-appended street-segments whose deflection angle is the smallest. By contrast, the self[-random]-fit join principle chooses randomly. The random variant generally gives the best *"fit"* [2].

Given a Boltzmann-mesoscopic system, let us denote by $\Pr(\Omega)$ the distribution of the numbers of configurations Ω of its mesoscopic objects o. If the average amount of surprisal $\sum_{\Omega} \Pr(\Omega) \ln \Omega$ is preserved, $\Pr(\Omega)$ most plausibly follows a Pareto distribution $\Pr(\Omega) \propto \Omega^{-\lambda}$ by virtue of Jaynes's Maximum Entropy principle [3, 4]. Assuming that an information network is such a system, the probability p_μ of its state μ yields

$$p_\mu \propto \prod_{o_\mu \in \{r_\mu, j_\mu\}} \Omega_{o_\mu}^{-\lambda} = \mathrm{e}^{-\lambda S_\mu} \quad \text{with} \quad S_\mu = \sum_{o_\mu \in \{r_\mu, j_\mu\}} \ln \Omega_{o_\mu} \tag{1}$$

the amount of surprisal in state μ; the product (so the sum) runs over the roads r_μ and junctions j_μ of state μ. This state probability depends only on the actual state of the information network. So, for a given self-organized urban street network, we can generate *Markov chains* [6] of information networks whose valence distribution reaches a Pareto distribution as equilibrium.

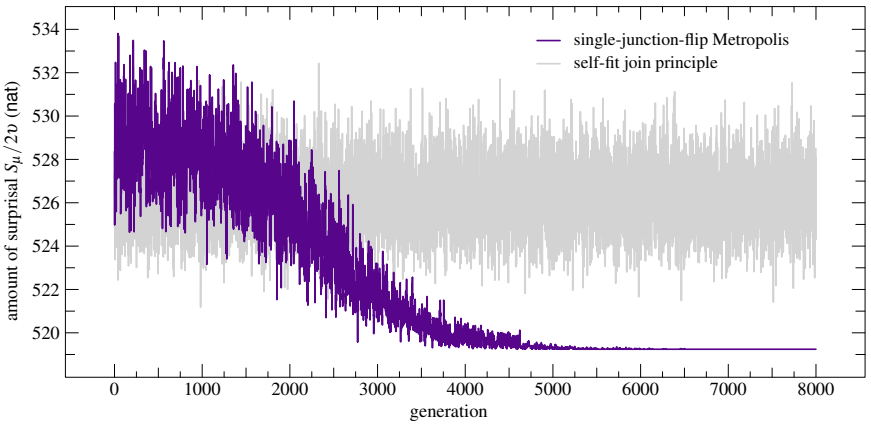


Fig. 1. Typical single-junction-flip Metropolis run for Old Ahmedabad (India): the foreground purple generation series plots a typical run starting from a self-fit state; the background light-grey generation series plots a typical random sequence of self-fit states with the same *modus operandi*. The simulations were run with a custom **C** code adopting and adapting typical techniques [6].

To achieve this, we must place on two conditions [6]. First, the *condition of ergodicity* assures that the Markov process can reach any state from any other one. The self-fit join principle readily inspires us the following *single-junction-flip* ergodic iteration: choose randomly a street-segment, then a direction towards one of its junctions,

then a new street-segment nominee, and finally recompose accordingly the arrangement of the chosen junction. Second, the *condition of detailed balance* assures both that every Markov chain comes to an equilibrium and that it is the probability distribution (1) which is effectively generated. An abundant literature exists on the subject [6]. For the sake of preliminary investigation, we adopt the Metropolis algorithm [6]. Thusly our *acceptance ratio* $A(\mu \to \nu)$ to accept a new state ν from state μ writes

$$A(\mu \to \nu) = \begin{cases} \mathrm{e}^{-\lambda(S_\nu - S_\mu)} & \text{if } S_\nu - S_\mu > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

In fact, we adapt the *single-spin-flip* variant of the Metropolis algorithm [6] since we use the single-junction-flip dynamics to generate new states. We refer to the literature to elaborate more sophisticated variants [6].

For early investigations, we may reduce urban street networks to their roads only so that the amount of surprisal S_μ in state μ simplifies [3, 4] to

$$S_\mu = 2\upsilon \sum_{r_\mu} \ln n_{r_\mu} \quad \text{since} \quad \Omega_{r_\mu} \propto n_{r_\mu}^{2\upsilon} \tag{3}$$

where υ is the *number of vital connections* for roads [3, 4]; the sum runs over the roads r_μ of state μ. We denotes by n_r the number of junctions of road r.

3 Results and Discussion

Figure 1 shows a typical single-junction-flip Metropolis run for the urban street network of Old Ahmedabad. Besides validating our approach, our runs give two precious indications. First, the actual convergence of typical runs confirms that information networks of Old Ahmedabad plausibly follow a scale-free coherence. Second, the convergent information network of least surprisal appear mostly unreachable through the self-fit join principle. Both encourage to reinforce the similitude with Ising-like models [6].

In future works, we expect to generate among self-organized information networks different surprisal equilibria. Ultimately, this may bring us a thermodynamic-like toolbox [6] to investigate and understand the geometrical rules and the social dynamics that actually govern urban street networks and, by extension, cities.

References

1. Rosvall, M., Trusina, A., Minnhagen, P., Sneppen, K.: Networks and cities: An information perspective. Phys. Rev. Lett. 94(2), 028701 (2005)
2. Jiang, B., Zhao, S., Yin, J.: Self-organized natural roads for predicting traffic flow: A sensitivity study. J. Stat. Mech. Theor. Exp. 2008(7), P07008 (2008)
3. Benoit, J., Jabari, S.: On the perturbation of self-organized urban street networks. Appl. Netw. Sci. 4, 49 (2019)
4. Benoit, J., Jabari, S.: Structure entropy, self-organization and power laws in urban street networks (Jan 2019), https://arxiv.org/abs/1902.07663
5. Alexander, C.: A city is not a tree. Arch. Forum 122(1+2), 58–62 (1965)
6. Newman, M.E.J., Barkema, G.T.: Monte Carlo Methods in Statistical Physics. Oxford University Press, Oxford (1999)

Consensus Partitioning in a Water Distribution Network based on Substance Propagations

Nicolas Cheifetz[1], Oussama Ennouri[1], Pierre Mandel[1],
Cédric Féliers[1], and Véronique Heim[2]

[1] Veolia Eau d'Ile-de-France, Le Vermont, 28, Boulevard de Pesaro, Nanterre F-92751, France,
{nicolas.cheifetz, oussama.ennouri}@veolia.com,
{pierre.mandel, cedric.feliers}@veolia.com
[2] Syndicat des Eaux d'Ile-de-France, 120 Boulevard Saint-Germain, Paris F-75006, France,
v.heim@sedif.com

Monitoring water quality continuously is essential to ensure the sanitary conditions for any Water Distribution System (WDS). For the last decades, Event Detection Systems (EDS) have been used to fulfill this task in Water Distribution Networks (WDN). Recent approaches are hydraulic model-based [17] and fully data-driven based on machine learning [13] or deep learning [4] techniques applied on water quality time series. The first hydraulic approach assumes the availability of some hydraulic modeling to simulate the water quality data which might be very costly and unreachable in practice for large WDS. The second one hardly succeeds in distinguishing between a normal operating event (exchanging water between District Meter Areas (DMA), emptying a tank,...) and a real pollutant event only by analyzing the collected time series. There is thus a prior need for the data-driven approach to identify some knowledge about operating configurations of the WDN and similar quality zones to characterize normal patterns. This paper introduces a new methodology to overcome this problem by partitioning the WDN graph in node clusters with similar water quality.

The proposed approach aims to summarize water quality zones in the WDN not only in space but also in time. As a next step, some EDS should trigger an alarm when an event cannot be explained by this water quality segmentation. The input data used in this study are water flow time series on pipes and concentration time series for all WDN nodes based on conservative tracer simulations. The normal operating configuration of the WDS is found as described in [3] based on water flows. Furthermore, two papers propose node clustering techniques with hydraulic path information for a similar problem to define quality zones [12], [14]. However, both of them apply a classical Kmeans (KM) on some impact feature matrix defined by concentration values of conservative tracers and have two drawbacks. First, the unaligned concentration time series from distinctive nodes can lead to similar deviations to KM centroids due to the Euclidian distance [12] which suggest to deal with longitudinal data and not the classical multivariate data. Second, the Kmeans algorithm can group nodes together with similar deviations on different tracer simulations which suggest to perform as many graph clusterings as hydraulic simulations. To solve these issues, we use respectively a Dynamic Time Warping (DTW) distance between the time series and a consensus clustering on all the tracer simulations. More precisely, we introduce the construction of a hybrid matrix that gathers together spatial and temporal information. The temporal attributes are summarized by the widely used DTW [15] as a distance between pairs of concen-

COMPLEX NETWORKS 2019

The 8^{th} International Conference on Complex Networks and Their Applications. 10 - 12 Dec., 2019, Lisbon, Portugal

tration time series. The spatial data are the shortest path in terms of water flow for each conservative tracer simulation from the source to all exposed nodes. Let us recall that, when looking for dense subgraphs, graph clustering is called community detection in graphs [7] and usually based on Newman-Girvan modularity optimization [10]. Our methodology is a node-attributed graph clustering based on a similarity matrix close to the approach [6] that was employing a Kmeans for the clustering step. Then, we use a consensus learning step to compute a single node-partition based on the S graph clusterings where S is the number of hydraulic simulations. Note that global quality functions, e.g. modularity, are known to have serious limits, and their optimization is often unable to detect clusters in realistic settings [10].

The present work extends a paper published at MARAMI'19 [5] with two major contributions: an Expectation-Maximization (EM) algorithm [9] is performed for each graph clustering and all the EM posterior probabilities are employed for an agglomerative hierarchical clustering as a consensus step (respectively a Kmeans and a graph coloring in the previous work). The similarity matrix is formulated in the same way in order to perform simultaneously a graph clustering on spatial and temporal data. A dedicated EM algorithm is used to produce a soft partition of the WDN for each tracer propagation. More precisely, a Gaussian parsimonious EM algorithm is implemented with diagonal covariance matrices [2] and the expectation step is modified to handle the non-exposed nodes where shortest paths can not be computed. Note that the initialization step is done by applying a Kmeans on the spatial vectors (shortest paths) to speed up the process which is found efficient in practice compared to the entire similarity matrix. The model selection is done by minimizing the BIC (Bayesian Information Criterion) criterion with a specific penalization. Concerning the consensus clustering on the S independent partitions, we employ a strategy varying the number of final groups [8] in the WDN graph. The graph coloring algorithm is the problem where adjacent vertices in a graph always must have distinct colors i.e cluster label in our work [16]. Such approach is found stable in practice but its final partition is static. In this paper, a specific hierarchical measure is formulated using the posterior probabilities of the Gaussian mixtures to evaluate the different arrangement of the clusters. A complete linkage is used and the number of groups is chosen by cutting a dendrogram at a chosen level. This explains also how the WDN clusters of similar water quality can merge or split according to the hierarchical measure. Note that the proposed segmentation methodology uses a reduced graph of the WDN (removing antennas and linear nodes) to simplify the problem. An algorithm of label propagation is run afterwards to cluster each unlabeled node of the global WDN graph. Finally, each cluster is characterized by a S-dimensional prototype concentration curve by applying the probabilistic generative method called Continuous Profile Model (CPM) [11]. This approach allows to define each cluster by its prototype curve in terms of source influence in space and time.

The methodology is illustrated on a large real-world network that belongs to the Syndicat des Eaux d'Ile-de-France (SEDIF). The SEDIF is a large association including 151 municipalities which produces about 780,000 m^3 of drinking water each day for about 4.6 million inhabitants of suburban Paris. This is the largest drinking WDN in France with about 8,700 km of pipes and its graph consists of an outer branched component (forest) and an inner one (core). The proposed approach is implemented on a

major part of the SEDIF network represented by a single calibrated model using the hydraulic modeling software Synergi^{TM} Water, with a graph of about 32k vertices and 40k edges. Based on this model calibrated with real-world data streams, the experimental results show a certain practicality in extracting spatio-temporal patterns from an operating WDS and exhibiting source influences of the WDN clusters. Such information about the dynamical topology of the WDN is valuable to enhance quality event detectors. Future work will investigate the comparison of various EDS in terms of change-point detection performance (e.g. [1], [13], [4]) based on the proposed methodology.

References

1. Ba, A., McKenna, S.A.: Water quality monitoring with online change-point detection methods. Journal of Hydroinformatics 17(1), 7–19 (2015)
2. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern recognition 28(5), 781–793 (1995)
3. Cheifetz, N., Kraiem, S., Mandel, P., Féliers, C., Heim, V.: Extracting temporal patterns for contamination event detection in a large water distribution system. In: 15th International Computing and Control for Water Industry conference (CCWI 2017). Sheffield, UK (Sep 2017)
4. Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C.: A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water Journal 16 (2019)
5. Ennouri, O., Cheifetz, N., Mandel, P., Féliers, C., Heim, V.: Graph clustering for quality event detectors in a large water distribution network. In: MARAMI 2019: The 10th Conference on Network Modeling and Analysis. Dijon, FR (Nov 2019)
6. Falih, I., Grozavu, N., Kanawati, R., Bennani, Y.: ANCA: Attributed Network Clustering Algorithm. In: International Conference on Complex Networks and their Applications. vol. 689, pp. 241–252. Springer (2017)
7. Fortunato, S.: Community detection in graphs. Physics reports 486(3-5), 75–174 (2010)
8. Fritsch, A., Ickstadt, K., et al.: Improved criteria for clustering based on the posterior similarity matrix. Bayesian analysis 4(2), 367–391 (2009)
9. Fruhwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of mixture analysis. Chapman and Hall/CRC (2019)
10. Lancichinetti, A., Fortunato, S.: Consensus clustering in complex networks. Scientific reports 2, 336 (2012)
11. Listgarten, J., Neal, R.M., Roweis, S.T., Emili, A.: Multiple alignment of continuous time series. In: Advances in Neural Information Processing Systems (NIPS). pp. 817–824 (2005)
12. Mandel, P., Maurel, M., Chenu, D.: Better understanding of water quality evolution in water distribution networks using data clustering. Water research 87, 69–78 (2015)
13. Muharemi, F., Logofătu, D., Leon, F.: Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication pp. 1–14 (2019)
14. Qin, T., Boccelli, D.L.: Grouping water-demand nodes by similarity among flow paths in water-distribution systems. Journal of Water Resources Planning and Management 143(8), 04017033 (2017)
15. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing 26(1), 43–49 (1978)
16. Voloshin, V.: Graph coloring: History, results and open problems. Alabama Journal of Mathematics, Spring/Fall (2009)
17. Yang, X., Boccelli, D.L.: Model-based event detection for contaminant warning systems. Journal of Water Resources Planning and Management 142(11), 04016048 (2016)

COMPLEX
NETWORKS

INSTITUTO GULBENKIAN DE CIÊNCIA

UNIVERSIDADE NOVA DE LISBOA

POLITÉCNICA

uc3m
Universidad Carlos III de Madrid

universidade de aveiro

UNIVERSITÉ DE LYON

TU Delft

UNIVERSITÀ DEGLI STUDI DI MILANO

Springer

frontiers

IOP Publishing

TRIBE

The 8th International Conference on Complex Networks and Their Applications
December 10 - 12, 2019 - Lisbon, Portugal

www.complexnetworks.org

9 782955 705032