On The Estimation of The Number of Communities for Sparse Networks

By Neil Hwang* and Jiarui Xu† and Shirshendu Chatterjee* and Sharmodeep Bhattacharyya†

City University of New York *
Oregon State University, Corvallis †

In most clustering algorithms, the number of communities, K, is a required in- put. Among the various approaches that have been proposed to estimate K, the non- parametric methods based on the spectrum of the associated Bethe Hessian matrices (H_ζ with parameter ζ) have garnered much popularity for its simplicity, computa- tional efficiency, and robustness to sparsity of data, which have been demonstrated in several empirical studies that have ensued. For certain heuristic choices of ζ, such methods have been recently shown to be consistent if the input network with N nodes is generated from the (semi-dense) stochastic block model, where all nodes have equal expected degree and the common expected degree is $\gg \log N$. In this paper, we obtain several finite sample results to show that if the input network is generated from either stochastic block models (SBM) or degree-corrected block models (DCBM) having pos- sible heterogeneity both in terms of the expected degrees of the nodes and the sizes of the communities, and if ζ is chosen from a certain interval, then the associated spec- tral methods based on H_ζ is consistent for estimating K not only for the semi-dense regime but also for the sub-logarithmic sparse regime, when the maximum (d) of the expected degrees of all nodes satisfies $1 \ll d \ll \log N$, under some mild condition on the extent of heterogeneity. We also propose a method to estimate the aforementioned interval empirically, which enables us to develop a consistent estimation procedure. We evaluate the performance of the resulting estimation procedure theoretically. The efficacy of our proposed method is demonstrated via extensive simulation studies and the application of our approach to a comprehensive collection of real-world network data arising in diverse areas of interest.

Keywords: Spectral Clustering, Sparse Networks, the Bethe Hessian Operator, Community Number Estimation, Stochastic Block Model, Degree-Corrected Block Model