Detection of an anomalous path in a noisy network

Shirshendu Chatterjee,
City College & Graduate Center, CUNY

Joint Work with Ofer Zeitouni

CUNY Probability Seminar, Feb 2018
The topic fits into broad framework of nonparametric detection.

Based on measurements about variables related through a graphical model, detect whether there is a sequence of connected nodes which exhibit a “peculiar behavior”.

Consider a graph-indexed process: a model problem for detecting whether or not there is a chain of connected nodes in a given network which exhibit an “unusual behavior”.

- Trace the existence of a polluter in a network of streams
- Detecting atypical gene behavior in a given gene network

Based on one realization of this process,

- can one reliably detect if there is a chain of nodes (hidden in the background noise) that stand out?
- How subtle a difference can one detect?
Suppose $G_n$ is a $n \times n$ two dimensional graph with node set $V_n$ and path set $P_n$. Paths in $P_n$ are nonintersecting and each has order $n$ many nodes. Suppose each node $v$ has a r.v. $X_v$ attached to it.

**Observable:** $(X_v, v \in V_n)$.

- **Null hypothesis** $H_0$: The random variables $\{X_v : v \in V_n\}$ are i.i.d. with common distribution $N(0, 1)$.
- **Alternate (signal) hypothesis** $H_{1,n}$: it is a composite hypothesis $\bigcup_{\pi \in P(G_n)} H_{1,\pi}$, where, under $H_{1,\pi}$, the random variables $\{X_v : v \in V_n\}$ are independent with

$$X_v \overset{d}{=} \begin{cases} N(\mu_n, 1) & \text{if } v \in \pi \\ N(0, 1) & \text{otherwise} \end{cases}$$

for some $\mu_n > 0$. 

S. Chatterjee
Detection of Anomalous Path
Let $\mathbb{P}_0$ and $\mathbb{P}_\pi$ be the probability distribution of $(X_v, v \in V_n)$ under the hypothesis $H_0$ and $H_{1,\pi}$.

**Question.** For which values of $\mu_n$, the probability distributions $\mathbb{P}_0$ and $\bigcup_{\pi \in \mathcal{P}_n} \mathbb{P}_\pi$ are distinguishable?

More precisely, consider the following.

- A test $T_n$ is a $\{0, 1\}$-valued function of $(X_v, v \in V_n)$
- $\{T_n = 1\} \iff$ “Accept signal hypothesis”, and $\{T_n = 0\} \iff$ “Accept null hypothesis”
- **Minimax risk:** For any test $T_n$, $\gamma(T_n) := \mathbb{P}(\text{Type I}) + \sup_{\pi \in \mathcal{P}_n} \mathbb{P}(\text{Type II}) = \mathbb{P}_0(T_n = 1) + \sup_{\pi \in \mathcal{P}_n} \mathbb{P}_\pi(T_n = 0)$.
- A test $T_n$ is
  - asymptotically powerful if $\gamma(T_n) \to 0$.
  - asymptotically powerless if $\gamma(T_n)$ is close to 1.

**Question.** What values of $\mu_n$ (signal per anomalous node) can be detected reliably?

S. Chatterjee

Detection of Anomalous Path
The GLRT would reject $H_0$ for large values of $M_n := \max_{\pi \in P_n} \sum_{v \in \pi} X_v$. Its performance is not optimal.
Consider the two dimensional graph with vertex set

\[ V_n := \{(i, j) : 0 \leq i \leq n - 1, |j| \leq i, \text{ and } i, j \text{ have same parity}\}, \]

and path set \( P_n \) consisting of directed paths starting at the origin.

**Theorem (Castro, Candes, Helgason, Zeitouni (2008))**

If \( \mu_n \sqrt{\log(n)} \rightarrow \infty \), then there is an asymptotically powerful test.
If \( \mu_n \log(n) \sqrt{\log \log(n)} \rightarrow 0 \), then all tests are asymptotically powerless.
The statistic separating the two hypothesis

- Consider a linear statistic \( \sum_{(i,j) \in V_n} w_{i,j} X_{(i,j)} \), where \((w_{i,j})_{(i,j) \in V_n}\) are weights.
- Choose the weights which maximizes the signal-noise ratio.
- Optimal weight \( w_{i,j} \) turns out to be proportional to \( 1/(i + 1) \).
- So the (weighted average) linear statistic separating the two hypothesis is
  \[
  \text{WAS} := \sum_{(i,j) \in V_n} \frac{X_{(i,j)}}{i + 1},
  \]
- \( \mathbb{E}_0(\text{WAS}) = 0 \), \( \mathbb{E}_\pi(\text{WAS}) = \mu_n \log(n) \), \( \text{Var}_0(\text{WAS}) = \text{Var}_\pi(\text{WAS}) = \log(n) \), so the signal-noise ratio is \( \mu_n \sqrt{\log(n)} \).
- So if \( \mu_n \sqrt{\log(n)} \to \infty \), then WAS separates the two hypothesis.
Consider the two dimensional graph with vertex set

\[ V_n := \{(i, j) : 0 \leq i \leq n-1, |j| \leq an+i, \text{ and } i, j \text{ have same parity}\}, \]

and path set \( P_n \) consisting of directed paths starting at the left hyperplane.

**Theorem (C. and Zeitouni (2017))**

If \( \mu_n \sqrt{\log(n)} \geq C \) for some large constant \( C \), then there is an asymptotically powerful test.

If \( \mu_n \log(n) \sqrt{\log \log(n)} \to 0 \), then all tests are asymptotically powerless.
The statistic separating the two hypothesis

- The WAS loses its power in this case.
- We develop a polynomial statistic (PS), which is a polynomial in \((X_{(i,j)}, (i, j) \in V_n)\).
- The PS is obtained inductively using a quadratic statistic as the building block.
- Call \((i, j) \leftrightarrow (i', j')\) if a path hitting \((i, j)\) can visit \((i', j')\) and vice versa.
- Define the quadratic forms

\[
Q_n := \sum_{(i, j), (i', j') \in V_n, (i, j) \leftrightarrow (i', j')} \frac{1}{|i - i'|} X_{(i, j)} X_{(i', j')}.
\]

- \(E_0(Q_n) = 0, E_{\pi}(Q_n) = \mu_n^2 n \log(n),\ Var_0(Q_n) = n^2 \log(n)\) and \(Var_{\pi}(Q_n) = (1 + o(1)) n^2 \log(n)\), so the signal-noise ratio is \(\mu_n^2 \sqrt{\log(n)}\).
- So if \(\mu_n \gg (\log(n))^{-1/4} \rightarrow \infty\), then \(Q_n\) separates the two hypothesis.
We partition the graph into disjoint squares and half-squares having side length $\sqrt{n}$, and consider

- the coarse grained graph, where each square (and half-square) represents a node,
- the coarse grained path on the above graph.

The random variable attached to a coarse grained node is the normalized quadratic form $Q_{\sqrt{n}}$ computed on the associated square (or half-square).
Thus, we get a renormalized version of the original problem, where $n$ is replaced by $\sqrt{n}$ and $\mu_n$ is replaced by $\mu_n^2 \sqrt{\log(n)} = c \mu_n^2 \sqrt{\log(n)}$ (the signal noise ratio for $Q_{\sqrt{n}}$).

If we compute a similar quadratic form of the quadratic forms corresponding to the squares (and thus use a polynomial of degree 4), then the new signal-noise ratio becomes

$$\asymp \left( \mu_n^2 \sqrt{\log(n)} \right)^2 \sqrt{\log(n)} = \mu_n^4 \log^{3/2}(n) \to \infty \text{ if } \mu_n \gg (\log(n))^{-3/8}.$$ 

Repeating the renormalization argument ‘few’ times, we get the upper bound for detection threshold to be $C/\sqrt{\log(n)}$. 
Suppose that the anomalous path $\pi$ follow a prior distribution $\Pi$ on the set of paths.

Let $L_n$ be the likelihood ratio: $L_n(X) := \frac{dP_\Pi}{dP_0}(X)$. Consider the test $1\{L_n \geq 1\}$.

It is well known that

$$\inf_{T_n} \gamma(T_n) = \gamma(1\{L_n \geq 1\}) = P_0(L_n \geq 1) + P_\Pi(L_n < 1) =: \gamma_n^*.$$ 

Clearly $E_0(L_n) = 1$. A standard calculation shows

$$\gamma_n^* = 1 - \frac{1}{2} E_0|L_n - 1|.$$

Need to find condition on $\mu_n$ such that $E_0|L_n - 1| \to 0$.

Instead, we use Cauchy-Schwartz inequality to have

$$\gamma_n^* \geq 1 - \frac{1}{2} \sqrt{E_0[(L_n - 1)^2]}$$

and find conditions on $\mu_n$ so that $E_0[(L_n - 1)^2] \to 0$. 

---

S. Chatterjee

Detection of Anomalous Path
• Recall that $L_n = \frac{d\Pi_n}{d\Pi_0}(X) = E_\Pi \exp(\sum_{v \in \Pi} \mu_n X_v - \frac{1}{2} \mu_n^2)$.

• $E_0[(L_n - 1)^2] = E_0(L_{m}^2) - 1$ and

$$E_0(L_n^2) = E_0 E_{\Pi_1 \times \Pi_2} \exp \left( \sum_{v \in \Pi_1} \mu_n X_v - \frac{1}{2} \mu_n^2 + \sum_{v \in \Pi_2} \mu_n X_v - \frac{1}{2} \mu_n^2 \right)$$

$$= E_{\Pi_1 \times \Pi_2} \left[ \prod_{v \in \Pi_1 \cap \Pi_2} E_0 e^{2\mu_n X_v - \mu_n^2} \prod_{v \in \Pi_1 \triangle \Pi_2} E_0 e^{\mu_n X_v - \frac{1}{2} \mu_n^2} \right]$$

$$= E_{\Pi_1 \times \Pi_2} e^{\mu_n^2 N_n},$$

where $N_n$ is the number of intersections between two independent walks having prior $\Pi$. 
The strategy is to construct a prior on the family of paths with a low predictability profile, that is, a process whose location in the future is hard to predict from its current state and history.

**Predictability profile** (Benjamini, Pemantle and Peres, 1998) of a stochastic process \((S_t)_{t \geq 0}\) is

\[
PRE_S(k) := \sup_{x, \text{history}} P(S_{t+k} = x \mid S_0, S_1, \ldots, S_t), \quad k \in \mathbb{N}.
\]

**Theorem (Haggstrom & Mossel (1998) improving Benjamini et al)**

If \((f_k)_{k \geq 1}\) is a decreasing and positive sequence such that

\[
\sum_{k \geq 1} \frac{f_k}{k} < \infty,
\]

then there exists a nearest-neighbor walk \((S_t)_{t \geq 0}\) starting at \(S_0 = 0\) satisfying \(PRE_S(k) \leq \frac{C}{kf_k}\).

Hoffman (1998) proved if \((f_k)_{k \geq 1}\) is a decreasing positive sequence with

\[
\sum_{k \geq 1} \frac{f_k}{k} = \infty,
\]

then the above predictability profile is impossible to achieve.
The number of intersections between two independent nearest-neighbor walks drawn from a prior with low predictability profile is “small”, namely it has exponential tails.

Lemma

For a walk \((S_t)_{0 \leq t \leq n-1}\) with finite number of steps, if

\[
\sum_{1 \leq k < n/B} \text{PRE}_{S}(kB) \leq \theta < 1 \text{ for some } B,
\]

then for any (possibly deterministic) sequence \((v_t)_{0 \leq t \leq n-1}\),

\[
P(|S \cap v| > k) \leq B \cdot \theta^{k/B}.
\]

Using this estimate and some work,

\[
\mathbb{E}_{\Pi_1 \times \Pi_2} e^{\mu_n^2 N_n} \lesssim e^{\mu_n^2 \log^2(n) \log \log(n)} \rightarrow 1 \text{ if } \mu_n \ll [\log(n) \sqrt{\log \log(n)}]^{-1}.
\]

This will imply \(\lim \inf \gamma_n^* \geq 1\). So reliable detection is impossible.
Our algorithm also applies to the ‘known initial location’ case.

Our algorithm also applies to the case where the paths are not necessarily directed.

The detection threshold vanishes as $n$ grows.

*Other distribution.* Similar results are available for other distributions from exponential family.

*Other graph.* The detection threshold is sensitive to the underlying graph. The threshold is nonvanishing when the graph has a tree structure or has high dimension.
The anomalous subset may be evolving with time.

- It may be associated to the unfolding of a stochastic process.
- It may be associated to or controlled by an underlying “particle system” running on the network.
Thank you