Inference in time series of graphs using locality statistics

Heng Wang, Minh Tang, Youngser Park, Carey E. Priebe

Johns Hopkins University
Department of Applied Mathematics and Statistics
Baltimore, Maryland 21218-2682 USA

hwang82@jhu.edu, mtang10@jhu.edu, youngser@jhu.edu, cep@jhu.edu

Abstract

We formulate change-point detection in a time series of graphs as a hypothesis testing problem in terms of Stochastic Block Model time series. We analyze two classes of scan statistics by deriving the limiting properties and power characteristics of the competing scan statistics.

1. Change-Point detection in Stochastic Block Model formulation

Given a time series of graphs $G_t = (V, E_t)$, where the vertex set $V = [n] = \{1, \cdots, n\}$ is fixed throughout, an important inference task in time series analysis is to identify, from $\{G_t\}$, excessive communication activities in a subregion of a dynamic network.
Statistically speaking, we want to test, for a given $t \in \mathbb{N}$, the null hypothesis H_0 that t is not a change-point against the alternative hypothesis H_A that t is a change-point.
We say that t^* is a change-point for $\{G_t\}$ if there exists distinct choices of matrices $\mathbf{P}^0$, $\mathbf{P}^A$ independent of t such that

$$H_0 : G_t \sim \mathrm{SBM}(\mathbf{P}^0, \{[n_i]\}) \text{ for all } t, \quad H_A : G_t \sim \begin{cases} \mathrm{SBM}(\mathbf{P}^0, \{[n_i]\}) & \text{for } t \le t^* - 1 \\ \mathrm{SBM}(\mathbf{P}^A, \{[n_i]\}) & \text{for } t \ge t^* \end{cases}$$

where $\mathrm{SBM}(\mathbf{P}, \{[n_i]\})$ denotes the stochastic blockmodel of [1], with block connectivity probabilities $\mathbf{P}$ and unknown block memberships $\{[n_i]\}$ In each block $[n_i]$, vertices follow the same probabilistic behavior and $\mathbf{P}$ is a $B \times B$ symmetric matrix where $\mathbf{P}_{j,k}$ denotes the block connectivity probability between blocks j and k
In this work, to illustrate a subset of vertices with chatter anomalous behavior in an otherwise stationary setting, we are concerned about is of the particular form for some $\delta > 0$,

$$\mathbf{P}^0 = \begin{pmatrix} p & p & \cdots & \cdots & p \\ p & h_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & h_{B-1} & p \\ p & \cdots & \cdots & p & p \end{pmatrix}, \mathbf{P}^A = \begin{pmatrix} p & p & \cdots & \cdots & p \\ p & h_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & h_{B-1} & p \\ p & \cdots & \cdots & p & p+\delta \end{pmatrix}.$$

The case where $h_2 > p, \ldots, h_{B-1} > p$ is of interest because we can consider each of the $[n_i]$ as representing a "chatty" group for time $t \le t^* - 1$, and at t^*, the previously non-chatty group $[n_B]$ becomes more chatty. The detection of this transition for the vertices in $[n_B]$ is one of the main reasons behind the locality statistics that will be explored.

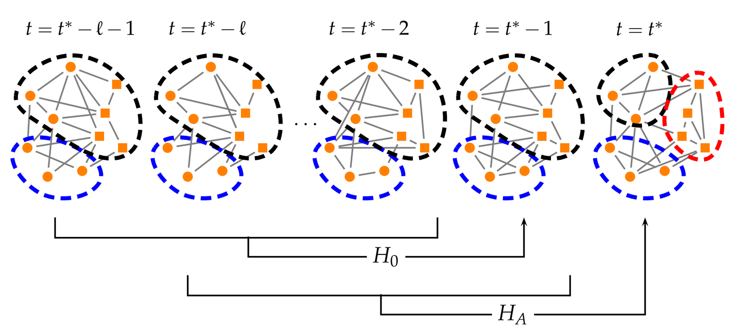


Figure 1: *Notional depiction of a 3-block time series of graphs in which the anomaly occurs at time t^*, a subset of vertices exhibits a change in behavior. When testing for change at time t^*, the recent past graphs $G_t, G_{t-1}, \ldots$ are used to standardize the invariants.*

2. Locality Statistics and Graph Invariants

2.1 Two locality statistics

Let $N_k[v; G] = \{u \in V : d(u, v) \le k\}$ and $\Omega(V', G)$ denote the subgraph of G induced by V'.
We now define two different but related locality statistics on $\{G_t\}$. For a given t, let $\Psi_{t;k}(v)$, introduced in [2], be defined for all $k \ge 1$ and $v \in V$ by

$$\Psi_{t;k}(v) = |E(\Omega(N_k(v; G_t); G_t))|. \tag{1}$$

$\Psi_{t;k}(v)$ counts the number of edges in the subgraph of G_t induced by $N_k(v; G_t)$.
Let t and t' be given, with $t' \le t$. Now define $\Phi_{t,t';k}(v)$, introduced in [3], for all $k \ge 1$ and $v \in V$ by

$$\Phi_{t,t';k}(v) = |E(\Omega(N_k(v; G_t); G_{t'}))|. \tag{2}$$

$\Phi_{t,t';k}(v)$ counts the number of edges in the subgraph of $G_{t'}$ induced by $N_k(v; G_t)$. Through this measure, a community structure shift of v can be captured even when the connectivity level of v remains unchanged across time.

2.2 Temporally-normalized statistics

Let $J_{t,t';k}$ be either the locality statistic $\Psi_{t';k}$ or $\Phi_{t,t';k}$, where for ease of exposition the index t is a dummy index when $J_{t,t';k} = \Psi_{t';k}$. With the purpose of determining whether t is a change-point, we now define two normalized statistics for $J_{t,t';k}$, a vertex-dependent normalization and a temporal normalization. These normalizations and their use in the change-point detection problem are depicted in Figure 1.
For a given integer $\tau \ge 0$ and $v \in V$, we define the vertex-dependent normalization $\widetilde{J}_{t,\tau;k}(v)$ of $J_{t,t';k}(v)$ by

$$\widetilde{J}_{t,\tau,k}(v) = (J_{t,t;k}(v) - \hat{\mu}_{t;\tau,k}(v))/\hat{\sigma}_{t;\tau,k} \tag{3}$$

where

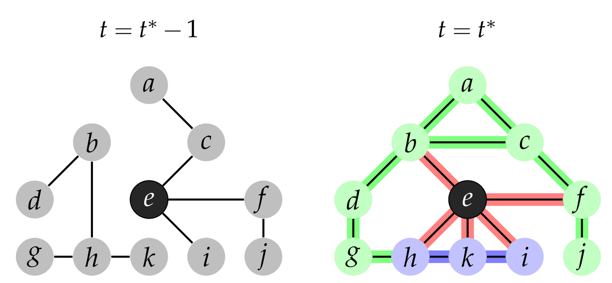
$$\hat{\mu}_{t;\tau,k}(v) = \frac{1}{\tau}\sum_{s=1}^{\tau} J_{t,t-s;k}(v), \quad \hat{\sigma}_{t;\tau,k}(v) = \mathbb{1}_{\{\tau\le1\}} + \mathbb{1}_{\{\tau>1\}}\sqrt{\frac{1}{\tau-1}\sum_{s=1}^{\tau}(J_{t,t-s;k}(v) - \hat{\mu}_{t;\tau,k}(v))^2}.$$

We then consider the maximum of these vertex-dependent normalizations $M_{\tau,k}(t) = \max_v(\widetilde{J}_{t,\tau;k}(v))$ and refer to $M_{\tau,k}(t)$ as the standardized scan statistics.
Finally, for a given integer $l \ge 0$, we define the temporal normalization of $M_{\tau,k}(t)$ by

$$S_{\tau,\ell,k}(t) = (M_{\tau,k}(t) - \tilde{\mu}_{\tau,\ell,k}(t))/\tilde{\sigma}_{\tau,\ell,k}(t) \tag{4}$$

where

$$\tilde{\mu}_{\tau,\ell,k}(t) = \frac{1}{\ell}\sum_{s=1}^{\ell} M_{\tau,k}(t-s), \quad \tilde{\sigma}_{\tau,\ell,k}(t) = \mathbb{1}_{\{\tau\le1\}} + \mathbb{1}_{\{\tau>1\}}\sqrt{\frac{1}{\ell-1}\sum_{s=1}^{\ell}(M_{\tau,k}(t-s) - \tilde{\mu}_{\tau,\ell,k}(t))^2}.$$



	$J^k_{t^*,t^*}(e)$		$\widehat{\mu}^k_{t^*,\tau}(e)$		$\widetilde{J}^k_{t^*,\tau}(e)$	
	$k=0$	$k=1$	$k=0$	$k=1$	$k=0$	$k=1$
$\Psi_{t';k}$	5	7	3	3	2	4
$\Phi_{t,t';k}$	5	7	2	4	3	3

Figure 2: *An example to differentiate the calculation of $\widetilde{J}_{t^*;\tau,k}(v)$ with varying underlying statistics ($\Psi_{t;k}$ or $\Phi_{t,t';k}$) and order distances ($k = 0$ or $k = 1$).*

3. Limiting Theory and Experiment

Theorem Under both H_0 and H_A, the limiting $S_{1,0,0}(t; \Psi), S_{1,0,0}(t; \Phi), S_{1,0,1}(t; \Psi)$ and $S_{1,0,1}(t; \Phi)$ are the maxima of random variables which, under proper normalizations, follow a standard Gumbel $\mathcal{G}(0, 1)$ distribution in the limit.
Corollary Let $\beta.$ be the power of the test statistic $S_{1,0,k}(t; \cdot)$ for $t = t^*$. As $n \to \infty$, neither $\beta.$ dominates when $k = 0$ and $\beta_\Phi \ge \beta_\Psi$ when $k = 1$.

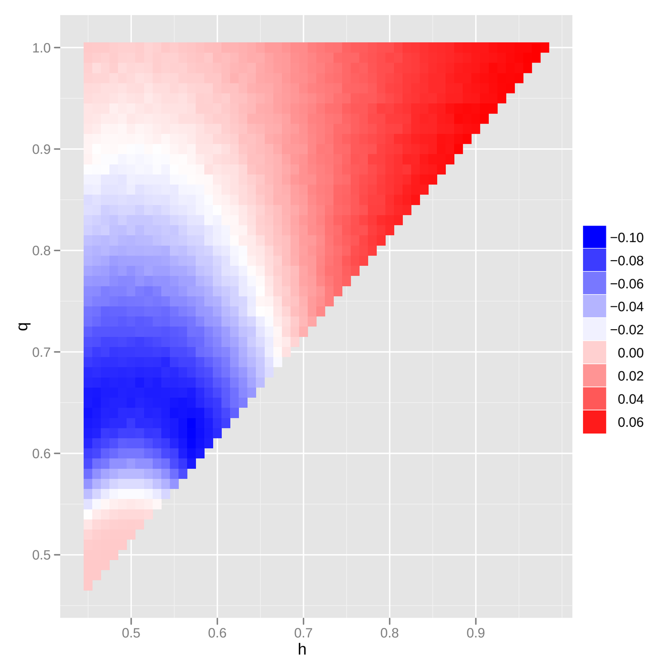


Figure 3: *A comparison, using the limiting properties of $S_{1,0,0}(t; \Psi)$ and $S_{1,0,0}(t; \Phi)$, of $\beta_\Psi - \beta_\Phi$ for different null and alternative hypotheses pairs as parametrized by h and $q(= p + \delta)$. The blue-colored region correspond to values of h and $q(= p + \delta)$ for which $\beta_\Psi < \beta_\Phi$ while the red-colored region correspond to values of h and $p + \delta$ with $\beta_\Psi > \beta_\Phi$.*

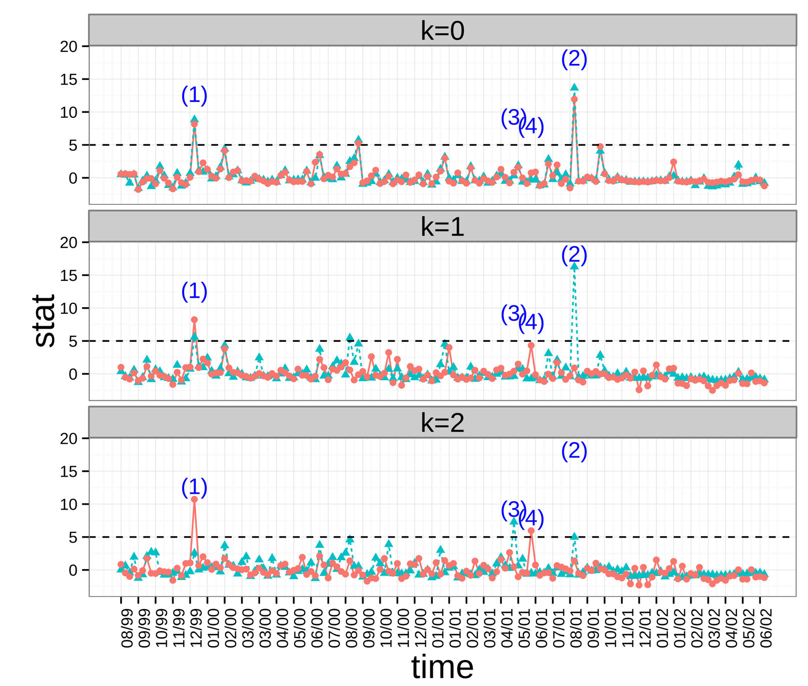


Figure 4: *$S_{\tau,\ell,k}(t; \Psi)$(sea green) and $S_{\tau,\ell,k}(t; \Phi)$(orange), the temporally-normalized standardized scan statistics using $\tau = \ell = 20$ with varing k, in time series of Enron email-graphs from August 1999 to June 2002. In the case $k = 0$, both $S_{20,20,0}(t; \Psi)$ and $S_{20,20,0}(t; \Phi)$ show detections ($S_{\tau,\ell,k}(t; \cdot) > 5$) at (1) and (2); in the case $k = 1$, both $S_{20,20,1}(t; \Psi)$ and $S_{20,20,1}(t; \Phi)$ show detections at (1), $S_{20,20,1}(t; \Phi)$ also indicates an anomaly at (2); in the case $k = 2$, $S_{20,20,2}(t; \Psi)$ detects anomalies at (2) and (3) but $S_{20,20,2}(t; \Phi)$ captures anomalies at (1) and (4). Detailed analyses on each observation are provided in [4].*

4. Future Work

Locality statistics based on Ψ can be readily computed in a real-time streaming data environment, in contrast to those based on Φ. Thus, discovering approximations of locality statistics based on Φ which simultaneously maintain better power characteristics and are amenable to streaming graphs is of interest.

References

[1] P. W. Holland, K. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, pp. 109–137, 1983.

[2] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, "Scan statistics on Enron graphs," *Computational and Mathematical Organization Theory*, vol. 11, pp. 229–247, 2005.

[3] X. Wan, J. Janssen, N. Kalyaniwalla, and E. Milios, "Statistical analysis of dynamic graphs," in *Proceedings of AISB06: Adapation in Artificial and Biological Systems*, 2006, pp. 176–179.

[4] H. Wang, M. Tang, Y. Park, and C. E. Priebe, "Locality statistics for anomaly detection in time series of graphs," *IEEE Transactions on Signal Processing*, 2013, accepted for publication, http://arxiv.org/abs/1306.0267.

This work is supported in part by the Johns Hopkins University Human Language Technology Center of Excellence and the XDATA program of the Defense Advanced Research Projects Agency (DARPA). Relevant manuscripts available at `http://www.cis.jhu.edu/~parky/CEP-Publications/journal.html`.