Inference in time series of graphs using locality statistics

Heng Wang* Minh Tang † Carey Priebe ‡
Dept. of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD, USA
hwang82@jhu.edu* mtang10@jhu.edu† cep@jhu.edu‡

Youngser Park§
Center of Imaging Science
Johns Hopkins University
Baltimore, MD, USA
youngser@jhu.edu§

***Abstract*—The ability to detect change-points in a dynamic network or a time series of graphs is an increasingly important task in many applications of the emerging discipline of graph signal processing. This paper formulates change-point detection as a hypothesis testing problem in terms of Stochastic Block Model time series. We analyze two classes of scan statistics, based on distinct underlying locality statistics presented in the literature. Our main contribution is the derivation of the limiting distributions and power characteristics of the competing scan statistics. Performance is compared theoretically, on synthetic data, and on the Enron email corpus. We demonstrate that both statistics are admissible in one simple setting, while one of the statistics is inadmissible in a second setting.**

***Index Terms*—Dynamic network, Anomaly Detection, Scan statistics, Hypothesis testing, Random Graphs**

I. Introduction

Dynamic network data are often readily observed, with vertices denoting entities and time evolving edges signifying relationships between entities, and thus considered as a time series of graphs which is a natural framework for investigation. An anomalous signal is broadly interpreted as constituting a deviation from some normal network pattern while a change-point is the time-window during which the anomaly appears.

In this paper, we approach the dynamic anomaly detection problem through the use of locality-based scan statistics. Scan statistics are commonly used in signal processing to detect a local signal in an instantiation of some random field [1]. The idea is to scan over a small time or spatial window of the data and calculate some locality statistic for each window. The maximum of these locality statistics is known as the scan statistic. Large values of the scan statistic suggests existence of nonhomogeneity, for example, a local region with significantly excessive communications. Under some homogeneity hypothesis, change-point detection can then be reduced to statistical hypotheses testing (c.f. § II) using scan statistics.

Specifically, we identify excessive communication activity in a subregion of a dynamic network by employing scan statistics $S_{\tau,\ell,k}(t;\cdot)$ with τ, ℓ, k defined in (c.f. § III).

We utilize two locality statistics, Ψ and Φ, building on [2] and [3] respectively. Ψ is introduced in [2] to detect the emergence of local excessive activities in time series of Enron graphs. Φ is proposed in [3] to detect communication pattern changes in their department email network. However, all these previous works are only experiment-oriented. Under the assumption that the time series of graphs is stationary before the change-point, this paper demonstrates that the limiting distributions of $S_{\tau,\ell,k}(t;\Psi)$ and $S_{\tau,\ell,k}(t;\Phi)$ are statistical multinomial mixtures of Gumbel distributions in representative case $\tau = 1, \ell = 0$. Through these limiting distributions, comparative power analysis between $S_{\tau,\ell,k}(t;\Psi)$ and $S_{\tau,\ell,k}(t;\Phi)$ is performed. We demonstrate that both Ψ and Φ are admissible if $k = 0$, while Ψ is inadmissible if $k = 1$.

A. Notation

In this paper, we consider only undirected and unweighted graphs without self-loops. Generally, a graph is denoted by G, with vertex set $V = V(G)$ and edge set $E = E(G)$. The number of vertices of a graph is usually denoted by n. For a graph G on n vertices, the vertex set is usually taken to correspond to the set $[n] = \{1, 2, \ldots, n\}$. In our subsequent discussion, we might also partition V into subsets, or blocks. If V is partitioned into B blocks of size $n_1, n_2, \ldots, n_B$ vertices, then, with a slight abuse of notation, we shall denote by $[n_i]$ the vertices in block i.

Let G be a graph. For any $u, v \in V$, we write $u \sim v$ if there exists an edge between u and v in G. We write $d(u, v)$ for the shortest path distance between u and v in G. For $v \in V$, we denote by $N_k[v; G]$ the set of vertices u at distance at most k from v, i.e., $N_k[v; G] = \{u \in V : d(u, v) \leq k\}$. For $V' \subset V$, $\Omega(V', G)$ is the subgraph of G induced by V'. Thus, $\Omega(N_k[v; G], G)$ is the subgraph of G induced by vertices at distance at most k from v.

II. Change-point detection problem in Stochastic Block model formulation

An important inference task in time series analysis is to infer, from $\{G_t\}$, if there exists anomalous activities, e.g., excessive phone calls among a subgroup in the 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. The following formulation is a reasonable and sufficiently general way to form the basis of our discussion.

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

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

where $\mathrm{SBM}(\mathbf{P}, \{[n_i]\})$ denotes the stochastic blockmodel of [4], with block connectivity probabilities $\mathbf{P}$ and unknown block memberships $\{[n_i]\}$. Specifically, V is partitioned into B distinct blocks $[n_1], \ldots, [n_B]$. 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 contrast, the null hypothesis, i.e. the nonexistence of change-point, is

$$H_0 : G_t \sim \mathrm{SBM}(\mathbf{P}^0, \{[n_i]\}) \text{ for all } t.$$

In the following, we discuss a specific form for $\mathbf{P}^0$ and $\mathbf{P}^A$, illustrating a subset of vertices with altered communication behavior in an otherwise stationary setting.

The change parameters $(t^*, \{[n_i]\}, \mathbf{P}^0, \mathbf{P}^A)$ we are concerned about is of the form, for some $\delta > 0$,

$$\mathbf{P}^0 = \mathbf{P} + diag(0, h_2 - p, \ldots, h_{B-1} - p, 0), \tag{1}$$

$$\mathbf{P}^A = \mathbf{P}^0 + diag(0, \ldots, 0, \delta) \tag{2}$$

where $\mathbf{P}$ is a matrix that every element is p and $n_1, n_2, \ldots, n_B$ being of size $(n_1, n_2, \ldots, n_B) = (\Theta(n), O(n), \ldots, O(n))$. For this form of $\mathbf{P}^0$ and $\mathbf{P}^A$, before the change-point, each of the blocks $i = 2$ up to $B-1$ have self-connectivity probability h_i. 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 $t \leq t^* - 1$, and at t^*, the previously non-chatty group $[n_B]$ becomes more chatty. See Fig. 1 for a notional illustration of $\mathbf{P}^0$ and $\mathbf{P}^A$ for the case of $B = 3$ blocks. The detection of this transition for the vertices in $[n_B]$ is one of the main reasons behind the locality statistics that will be introduced in § III.

Fig. 1. Notional depiction of $\mathbf{P}^0$ and corresponding $\mathbf{P}^A$. $\mathbf{P}^0$: all vertices connect with probability p except that the self-connectivity probability of $[n_2]$ is h; $\mathbf{P}^A$: the self-connectivity probability of $[n_3]$ transitions from p to $p+\delta$ while $[n_2]$ retains its previous behavior.

III. Locality statistics for change-point detection in time series of graphs

A. Two locality statistics

Suppose we are given a time series of graphs $\{G_t\}_{t \geq 1}$ where G_t are constructed on the same vertex set V. We now define two different but related locality statistics on $\{G_t\}$. For a given t, let $\Psi_{t;k}(v)$ be defined for all $k \geq 1$ and $v \in V$ by

$$\Psi_{t;k}(v) = |E(\Omega(N_k(v; G_t); G_t))|. \tag{3}$$

$\Psi_{t;k}(v)$ counts the number of edges in the subgraph of G_t induced by $N_k(v; G_t)$, the set of vertices u at a distance at most k from v in G_t. In a slight abuse of notation, we let $\Psi_{t;0}(v)$ denote the degree of v in G_t. The statistic Ψ_t was introduced in [2]. [2] investigated the use of Ψ_t in analyzing the Enron data corpus.

Let t and t' be given, with $t' \leq t$. Now define $\Phi_{t,t';k}(v)$ for all $k \geq 1$ and $v \in V$ by

$$\Phi_{t,t';k}(v) = |E(\Omega(N_k(v; G_t); G_{t'}))|. \tag{4}$$

The statistic $\Phi_{t,t';k}(v)$ counts the number of edges in the subgraph of $G_{t'}$ induced by $N_k(v; G_t)$. Once again, with a slight abuse of notation, we let $\Phi_{t,t';0}(v)$ denote the degree of v in $G_t \cap G_{t'}$, where $G \cap G'$ for G and G' with $V(G) = V(G')$ denotes the graph $(V(G), E(G) \cap E(G'))$. The statistic $\Phi_{t,t';k}(v)$ is motivated by a statistic named the permanent window metric introduced in [3].

$\Phi_{t,t';k}(v)$ uses the community structure $N_k(v; G_t)$ at time t in its computation of the locality statistic at time $t' \leq t$.

B. Temporally-normalized statistics

Let $J_{t,t';k}$ be either the locality statistic $\Psi_{t';k}$ in Eq. (3) or $\Phi_{t,t';k}$ in Eq. (4), where for ease of exposition the index t is a dummy index when $J_{t,t';k} = \Psi_{t';k}$. 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 Fig. 2.

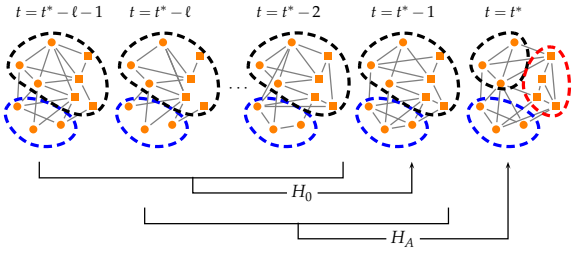

Fig. 2. Temporal standardization: when testing for change at time t, the recent past graphs $G_t, G_{t-1}, \ldots$ are used to standardize the invariants.

For a given integer $\tau \geq 0$ and $v \in V$, we define the vertex-dependent normalization $\widetilde{J}_{t,\tau;k}(v)$ of $J_{t,t';k}(v)$ by

$$\widetilde{J}_{t;\tau,k}(v) = \begin{cases} J_{t,t;k}(v) & \tau = 0 \\ J_{t,t;k}(v) - \hat{\mu}_{t;\tau,k}(v) & \tau = 1 \\ (J_{t,t;k}(v) - \hat{\mu}_{t;\tau,k}(v))/\hat{\sigma}_{t;\tau,k} & \tau > 1 \end{cases}, \tag{5}$$

where $\mu_{t;\tau,k}$ and $\sigma_{t;\tau,k}$ are defined as

$$\hat{\mu}_{t;\tau,k}(v) = \frac{1}{\tau} \sum_{s=1}^{\tau} J_{t,t-s;k}(v), \tag{6}$$

$$\hat{\sigma}_{t;\tau,k}(v) = \sqrt{\frac{1}{\tau - 1} \sum_{s=1}^{\tau} (J_{t,t-s;k}(v) - \hat{\mu}_{t;\tau,k}(v))^2}. \tag{7}$$

We then consider the maximum of these vertex-dependent temporal normalization for all $v \in V$, i.e., we define a $M_{\tau,k}(t)$ by

$$M_{\tau,k}(t) = \max_v (\widetilde{J}_{t,\tau;k}(v)). \tag{8}$$

We shall refer to $M_{\tau,0}(t)$ as the standardized max-degree and to $M_{\tau,1}$ as the standardized scan statistics. Finally, for a given integer $l \geq 0$, we define the temporal normalization of $M_{\tau,k}(t)$ by

$$S_{\tau,\ell,k}(t) = \begin{cases} M_{\tau,k}(t) & \ell = 0 \\ M_{\tau,k}(t) - \tilde{\mu}_{\tau,\ell,k}(t) & \ell = 1 \\ (M_{\tau,k}(t) - \tilde{\mu}_{\tau,\ell,k}(t))/\tilde{\sigma}_{\tau,\ell,k}(t) & \ell > 1 \end{cases}, \tag{9}$$

where $\tilde{\mu}_{\tau,\ell,k}$ and $\tilde{\sigma}_{\tau,\ell,k}$ are defined as

$$\tilde{\mu}_{\tau,\ell,k}(t) = \frac{1}{\ell} \sum_{s=1}^{\ell} M_{\tau,k}(t-s), \tag{10}$$

$$\tilde{\sigma}_{\tau,\ell,k}(t) = \sqrt{\frac{1}{\ell - 1} \sum_{s=1}^{\ell} (M_{\tau,k}(t-s) - \tilde{\mu}_{\tau,\ell,k}(t))^2}. \tag{11}$$

The statistics $S_{\tau,\ell,k}$ were defined to capture excessively increasing communications among a subset of vertices. We will use these $S_{\tau,\ell,k}$ as the test statistics for the change-point detection problem described in § II. For convenience of notation, since $S_{\tau,\ell,k}(t)$ is essentially a function of the $J_{t,t';k}$, we denote by $S_{\tau,\ell,k}(t; \Psi)$ and $S_{\tau,\ell,k}(t; \Phi)$ the $S_{\tau,\ell,k}(t)$ when the underlying statistic $J_{t,t';k}$ is $\Psi_{t',k}$ and $\Phi_{t,t';k}$, respectively.

IV. Power Estimates of $S_{\tau=1,\ell=0,k=0}(t; \cdot)$

In Section IV, we will derive the limiting distributions of $S_{\tau=1,\ell=0,k=0}(t; \cdot)$, showing that $S_{1,0,0}(t; \cdot)$ is a statistical multinomial mixture of Gumbel-distributed random variables. To clarify the notation, let $\mathcal{G}(\mu, \gamma)$ denote the Gumbel distribution with location

parameter μ and scale parameter γ. For theorems and propositions in Section IV and V, $S \xrightarrow{d} \sum_{i=1}^{B} \pi(n_i;\cdot)\mathcal{G}(\mu(n_i;\cdot),\gamma(n_i;\cdot))$ means that there exists $Z \sim Multinomial(1,\vec{\pi})$ such that $\frac{S-\mu(n_b;\cdot)}{\gamma(n_b;\cdot)} \xrightarrow{d} \mathcal{G}(0,1)$ given $Z = b$. Moreover, due to space restriction, detailed Gumbel parameters in the following Theorem 1 and Proposition 3 are provided in [5].

Theorem 1. *Let $\{G_t\}$ be a time series of random graphs according to the alternative H_A detailed in § II. In particular, $G_t \sim SBM(\mathbf{P}^0,\{[n_i]\}_{i=1}^B)$ for $t \le t^*-1$ and $G_t \sim SBM(\mathbf{P}^A,\{[n_i]\}_{i=1}^B)$ for $t \ge t^*$ with $\mathbf{P}^0$ and $\mathbf{P}^A$ being of the form in (1) and (2), respectively. Let $S_{1,0,0}(t;\cdot)$ denote the statistic $S_{\tau,l,k}(t;\cdot)$ with $\tau=1$, $l=0$, and $k=0$. Then as $n=\sum n_i \to \infty$, both $S_{1,0,0}(t;\Psi)$ and $S_{1,0,0}(t;\Phi)$ converge in distribution to a statistical multinomial mixture of Gumbel-distributed random variables i.e.,*

$$S_{1,0,0}(t;\cdot) \xrightarrow{d} \sum_{i=1}^{B} \pi_0(n_i;\cdot)\mathcal{G}(\mu_0(n_i;\cdot),\gamma_0(n_i;\cdot)) \quad t<t^*,$$

$$S_{1,0,0}(t;\cdot) \xrightarrow{d} \sum_{i=1}^{B} \pi_A(n_i;\cdot)\mathcal{G}(\mu_A(n_i;\cdot),\gamma_A(n_i;\cdot)) \quad t=t^*.$$

We note the following corollary to Theorem 1 for the case of $B=3$ blocks.

Corollary 2. *Assume the setting in Theorem 1 with $B=3$. Let $\alpha>0$ be given. Let $\beta_\cdot$ be the power of the test statistic $S_{1,0,0}(t;\cdot)$ for $t=t^*$ at significance level α. Then, as $(n_1,n_2,n_3)=(\Theta(n),O(n),O(n))$, β_Φ,β_Ψ and α have the following relationship:*

1) $n_3=o(\sqrt{n})$ *implies* $\beta_\Phi=\alpha,\beta_\Psi=\alpha$.
2) $n_3=\Omega(\sqrt{n})$ *implies* $\beta_\Psi>\alpha$.
3) $n_3=\Theta(\sqrt{n})=\Theta(n_2)$ *implies* $\beta_\Phi>\alpha$.
4) $n_3=\omega(\sqrt{n})=\Theta(n_2)$ *implies*

$$\beta_\Phi=\alpha \qquad \text{if } \lim_{n\to\infty}\frac{n_2(h(1-h)-p(1-p))}{n_3\delta(1-p)}>1,$$
$$\beta_\Phi>\alpha \qquad \text{if } \lim_{n\to\infty}\frac{n_2(h(1-h)-p(1-p))}{n_3\delta(1-p)}\le 1.$$

5) $n_3=\Omega(\sqrt{n})=\omega(n_2)$ *implies* $\beta_\Phi>\alpha$.
6) $n_3=\Omega(\sqrt{n})=o(n_2)$ *implies*

$$\beta_\Phi=\alpha \text{ if } h+p<1,$$
$$\beta_\Phi>\alpha \text{ if } h+p\ge 1.$$

From Corollary 2, an unanswered question is whether there exists a dominance between $S_{1,0,0}(t;\Phi)$ and $S_{1,0,0}(t;\Psi)$. By using Theorem 1, we now present an example to show that both statistics are admissible. Our setup is as follows. Let $p=0.43$. For each pair $(h,p+\delta)$ satisfying $p<h<1$ and $p<p+\delta<1$, we generate a null and alternative hypothesis pair H_0 and H_A according to the model in § II with $B=3$ blocks. $n=n_1+n_2+n_3=1000$ and n_1,n_2,n_3 are functions of n,h and δ ($n_2=n_3=c_{p,h,\delta}\sqrt{n\log n}$ where the constant $c_{p,h,\delta}$ is dependent on p,h and δ). In order to compare sensitivities of $S_{1,0,0}(t;\Psi)$ and $S_{1,0,0}(t;\Phi)$ in detection, we then calculate $\beta_\Psi-\beta_\Phi$ by deriving the limiting distributions of $S_{1,0,0}(t;\Psi)$ and $S_{1,0,0}(t;\Phi)$ using Theorem 1. The result is illustrated in Fig. 3 where we have plotted $\beta_\Psi-\beta_\Phi$ for different combinations of h and $q(=p+\delta)$. Fig. 3 indicates that the two statistics $S_{1,0,0}(\cdot;\Psi)$ and $S_{1,0,0}(\cdot;\Phi)$ are both admissible.

V. Power Estimates of $S_{\tau=1,\ell=0,k=1}(t;\cdot)$

In this section, we provide investigations of $S_{\tau,\ell,k}(t;\cdot)$ with larger scale parameter $k=1$ instead of $k=0$. We keep τ and ℓ the same as before, i.e., $\tau=1,\ell=0$. To make conclusions concise and presentable, firstly, we delve into the limiting distributions in the model presented in § II with number of blocks $B=3$.

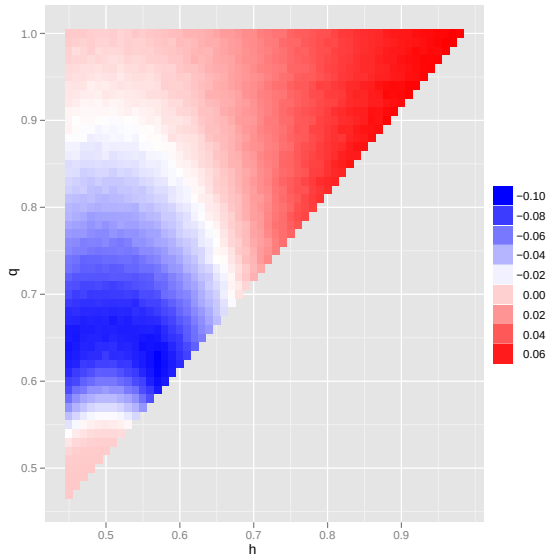

Fig. 3. A comparison, using the limiting distributions of $S_{1,0,0}(t;\Psi)$ and $S_{1,0,0}(t;\Phi)$, of $\beta_\Psi-\beta_\Phi$ for different null and alternative hypotheses pairs as parametrized by h and $q(=p+\delta)$. The blue-colored region correspond to values of h and $q(=p+\delta)$ for which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

Proposition 3. *Assume the same setting in Theorem 1 with $B=3$. As $(n_1,n_2,n_3)=(\Theta(n),o(n),o(n))$ and $n\to\infty$, $S_{1,0,1}(t;\Psi)$ converges in distribution to a statistical multinomial mixture of Gumbel-distributed random variables and so does $S_{1,0,1}(t;\Phi)$.*

Proposition 4. *In the model shown in Fig.1, Let $\alpha>0$ be given, $\beta'_\cdot$ be the power of the test statistic $S_{1,0,1}(t;\cdot)$ for $t=t^*$ at significance level α As $n\to\infty$, β'_Φ,β'_Ψ and α have the following relationship:*

1) $n_3=o(\sqrt{n})$ *implies* $\beta'_\Phi=\beta'_\Psi=\alpha$.
2) $n_3=\Omega(\sqrt{n})$ *implies* $\beta'_\Phi\ge\beta'_\Psi>\alpha$.

Consequently, Proposition 4 leads to the conclusion that the performance of $S_{1,0,1}(t;\Phi)$ dominates $S_{1,0,1}(t;\Psi)$ in the 3-block model. Moreover, this superiority can be generalized to the case with any given number of blocks $B\ge 3$. This is because each block $[n_i]$ with $1<i<B$ in B-blocks model follows a similar probabilistic behavior as block $[n_2]$ in 3-blocks model while both β'_Φ and β'_Ψ in B-blocks model can be characterized as a function of p,δ,n_B only. In other words, though $h_2>p,\ldots,h_{B-1}>p$, the "chatty" groups $[n_2],\ldots,[n_{B-1}]$ do not make any contribution on β'_Φ or β'_Ψ. Hence, the number of "chatty groups", namely $B-2$, is independent of the fact of dominance of $S_{1,0,1}(t;\Phi)$. Due to the superiority of $S_{1,0,1}(t;\Phi)$, only the limiting distribution of $S_{1,0,1}(t;\Phi)$ in the general B-block model is derived below.

Theorem 5. *Assume the same setting in Theorem 1. Let $S_{1,0,1}(t;\Phi)$ denote the statistic $S_{\tau,l,k}(t;\Phi)$ with $\tau=1$, $l=0$, and $k=1$. For a given $n\in\mathbb{N}$, let a_n and b_n be given by*

$$a_n=\sqrt{2\log n}\Big(1-\frac{\log\log n+\log 4\pi}{4\log n}\Big), b_n=\frac{1}{\sqrt{2\log n}}.$$

Then as $n=\sum n_i\to\infty$, $S_{1,0,1}(t;\Phi)$ converges in distribution to a statistical multinomial mixture of Gumbel-distributed random

variables. , i.e.,

$$S_{1,0,1}(t;\Phi) \xrightarrow{d} \sum_{i=1}^{B} \pi'_0(n_i;\Phi)\mathcal{G}(\mu'_0(n_i;\Phi),\gamma'_0(n_i;\Phi))\ t < t^*,$$

$$S_{1,0,1}(t;\Phi) \xrightarrow{d} \sum_{i=1}^{B} \pi'_A(n_i;\Phi)\mathcal{G}(\mu'_A(n_i;\Phi),\gamma'_A(n_i;\Phi))\ t = t^*,$$

where

$$\begin{aligned}
\eta(p) &= p^3(1-p)\\
\xi_0(n_i;\Phi) &= \mathbf{1}_{\{i\notin\{1,B\}\}} n_i(h_i(1-h_i) - p(1-p))\\
\mu'_0(n_i;\Phi) &= a_{n_i}\sqrt{Cn^2\eta(p)} + np(1-p) + \xi_0(n_i;\Phi)\\
\gamma'_0(n_i;\Phi) &= b_{n_i}\sqrt{Cn^2\eta(p)}
\end{aligned}$$

$$\begin{aligned}
\zeta(n_B,p,\delta,i) = \frac{\delta}{2}[&n_B^2(\mathbf{1}_{\{i\neq B\}}p^2 + \mathbf{1}_{\{i=B\}}(p+\delta)^2)\\
&+n_B(\mathbf{1}_{\{i\neq B\}}p(1-p) + \mathbf{1}_{\{i=B\}}(p+\delta)(1-p-\delta))]
\end{aligned}$$

$$\begin{aligned}
\mu'_A(n_i;\Phi) &= \mu'_0(n_i;\Phi) + \mathbf{1}_{\{i=B\}}n_B\delta(1-p) + \zeta(n_B,p,\delta,i)\\
\gamma'_A(n_i;\Phi) &= \gamma'_0(n_i;\Phi)
\end{aligned}$$

Corollary 6. *Assume the setting in Theorem 5. Let β' be the power of the test statistic $S_{1,0,1}(t;\cdot)$ for $t = t^*$. Then, as $(n_1, n_2, \ldots, n_B) = (\Theta(n), o(n), \ldots, o(n))$ and $n \to \infty$, $\beta'_\Phi \geq \beta'_\Psi$ and thus $S_{1,0,1}(t;\Psi)$ is inadmissible.*

VI. Experiment

We use the Enron email data used in [2] for this experiment. It consists of time series of graphs $\{G_t\}$ with $|V| = 184$ vertices and undirected edges for each week $t = 1, \ldots, 189$, where we draw a unweighted edge when vertex v sends at least one email to vertex w during a one week period.

Figure 4 depicts $S_{\tau,\ell,k}(t;\Psi)$ using dashed lines and $S_{\tau,\ell,k}(t;\Phi)$ using solid lines for a 20 week period from February 2001 through June 2001 (both $\tau = \ell = 20$ were used in [2]). As indicated in [2], detections are defined as weeks t such that $S_{\tau,\ell,k} > 5$. We observe from bottom Figure 4 that the second order scan statistic, i.e. $k = 2$, using $S_{\tau,\ell,k}(t;\Psi)$ indicates a clear anomaly at $t^* = 132$ in May 2001. For $S_{\tau,\ell,k}(t;\Phi)$, with the same detection condition, it is also apparent that there is a detection with max-degree ($k = 0$) at week $t^* = 132$ and is another one with the second order scan statistic at week $t^* = 136$ in June 2001.

Both detections using $S_{\tau,\ell,k}(t;\Phi)$, however, yield different v^*'s (`j.lavorato` and `m..scott`) from the one using $S_{\tau,\ell,k}(t;\Psi)$ (`k..allen`). This indicates that by using different locality statistic we can achieve different detections.

VII. Conclusion & Discussion

The simulation experiments indicate that the analytic power estimates, even when they are limited in scope, are useful in answering some important questions about the locality statistics. In particular, it was shown that Ψ and Φ are both admissible with respect to one another when $\tau = 1, \ell = 0, k = 0$. In addition, if $\tau = 1, \ell = 0, k = 1$, it is worthwhile to note that Ψ, compared with Φ, is inadmissible but computationally inexpensive. The locality statistics based on Ψ can be readily computed in a real-time streaming data environment, in contrast to those based on Φ. Thus, the adaption or approximation of locality statistics based on Φ for streaming environments is of interest. The investigations presented in this paper do not take into account attributes on the edges. The incorporation of edge attributes into the current paper is, however, straightforward. For example, [6] handles attributes by linear fusion, and many of the results there can be adapted to the current paper. In particular, one can define fused locality statistics for attributed graphs. Power estimates for these locality statistics can be derived in a similar manner to those presented in this paper. Other considerations, e.g., construction of fusion of graph invariants in [7] and corresponding optimal fusion parameters, can also be investigated.

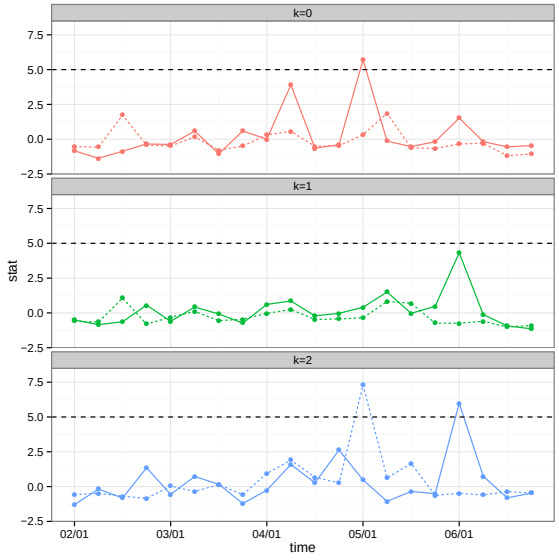

Fig. 4. $S_{\tau,\ell,k}(t)$, the temporally-normalized standardized scan statistics using $\tau = \ell = 20$, on zoomed-in time series of Enron e-mail graphs during a period of 20 weeks in 2001. Top: $k = 0$; Middle: $k = 1$; Bottom: $k = 2$. Dashed lines show a detection using $S_{\tau,\ell,k}(t;\Psi)$ (a standardized statistic $\widetilde{M}_{\tau,k}(t)$ which achieves a value greater than 5 standard deviations above its running mean) at week $t = 132$ in May 2001 for scale $k = 2$, but not for $k = 0$ or $k = 1$. Solid lines show detections using $S_{\tau,\ell,k}(t;\Phi)$ at week $t^* = 132$ in May 2001 for scale $k = 0$ and $t^* = 136$ in June 2001 for scale $k = 2$, both yield different v^*'s from the one in $S_{\tau,\ell,k}(t;\Psi)$.

Acknowledgment

This work was partially supported by Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.

References

[1] J. Glaz, J. Naus, and S. Wallenstein, *Scan Statistics*. Springer, 2001.
[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] P. W. Holland, K. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, pp. 109–137, 1983.
[5] H. Wang, M. Tang, Y. Park, and C. E. Priebe, "Locality statistics for anomaly detection in time series of graphs," 2013, arXiv preprint, http://arxiv.org/abs/1306.0267.
[6] M. Tang, Y. Park, N. H. Lee, and C. E. Priebe, "Attribute fusion in a latent process model for time series of graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1721–1732, April 2013.
[7] 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*, vol. 7, no. 1, pp. 67–75, Feb 2013.