Taylor & Francis
Taylor & Francis Group

Sequential Analysis
Design Methods and Applications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsqa20

Online score statistics for detecting clustered change in network point processes

Rui Zhang, Haoyun Wang & Yao Xie

To cite this article: Rui Zhang, Haoyun Wang & Yao Xie (2023) Online score statistics for detecting clustered change in network point processes, Sequential Analysis, 42:1, 70-89, DOI: 10.1080/07474946.2022.2164307

To link to this article: https://doi.org/10.1080/07474946.2022.2164307

SEQUENTIAL ANALYSIS
2023, VOL. 42, NO. 1, 70–89
https://doi.org/10.1080/07474946.2022.2164307

Taylor & Francis
Taylor & Francis Group

Online score statistics for detecting clustered change in network point processes

Rui Zhang, Haoyun Wang, and Yao Xie

School of Industrial and Systems Engineering (ISyE), Georgia Institute of Technology, Atlanta, Georgia, USA

ABSTRACT

We consider online monitoring of the network event data to detect local changes in a cluster when the affected data stream distribution shifts from one point process to another with different parameters. Specifically, we are interested in detecting a change point that causes a shift of the underlying data distribution that follows a multivariate Hawkes process with exponential decay temporal kernel, whereby the Hawkes process is considered to account for spatiotemporal correlation between observations. The proposed detection procedure is based on scan score statistics. We derive the asymptotic distribution of the statistic, which enables the self-normalizing property and facilitates the approximation of the instantaneous false alarm probability and the average run length. When detecting a change in the Hawkes process with nonvanishing self-excitation, the procedure does not require estimating the postchange network parameter while assuming the temporal decay parameter, which enjoys computational efficiency. We further present an efficient procedure to accurately determine the false discovery rate via importance sampling, as validated by numerical examples. Using simulated and real stock exchange data, we show the effectiveness of the proposed method in detecting change while enjoying computational efficiency.

ARTICLE HISTORY
Received 21 June 2022
Revised 18 November 2022
Accepted 11 December 2022

KEYWORDS
Change point detection; graph scanning statistics; score statistics

SUBJECT CLASSIFICATIONS
Primary 62L10; Secondary 62G10; 62G32

1. INTRODUCTION

Network Hawkes point processes recently became a popular model for sequential events data over networks, a widely encountered data type in modern applications (Hawkes 1971). Such data usually capture the temporal and spatial information of the events; that is, a sequence of event times, the corresponding event location, and additional information. A multivariate Hawkes process can model the influence of previous events on subsequent events; for instance, triggering the subsequent events or making them more likely to happen. Hawkes processes have been widely used in many areas such as finance (Hawkes 2018), social media (Rizoiu et al. 2017), epidemiology (Rizoiu et al. 2018), and seismology (Ogata 1998).

Change point detection is a fundamental problem in statistics, aiming to detect a transition in the distribution of the sequential data, which often represents a state transition

CONTACT Yao Xie yao.xie@isye.gatech.edu School of Industrial and Systems Engineering (ISyE), Georgia Institute of Technology, Atlanta, 765 Ferst Dr NW, GA 30332, USA.
Rui Zhang and Haoyun Wang contributed equally to this article.

(see, e.g., Moustakides 2008; Poor and Hadjiliadis 2008; Y. Xie and Siegmund 2013; Veeravalli and Banerjee 2014; Tartakovsky 2019; L. Xie et al. 2021). For example, in water quality monitoring, a change point can be due to a water contamination event (Chen, Kim, and Xie 2020); in public health, it may represent a disease outbreak (Christakis and Fowler 2010). The goal is to develop a procedure that can raise the alarm as soon as possible after the change point while controlling the false alarm rate.

Detecting change points for Hawkes processes is an important problem for monitoring large-scale networks using discrete events data. Online detection of change point detection in Hawkes processes is considered challenging due to the asynchronous nature of discrete events and temporal dependence, which is far from the traditional independent and identically distributed setting considered in change point detection literature. In H. Wang et al. (2022), the authors proposed an adapted version of the cumulative sum procedure to account for the dynamic behavior assuming known postchange parameters. When postchange parameters are unknown, a classic method is the generalized likelihood ratio (GLR) procedure, and Li et al. (2017) use the expectation-maximization type of algorithm to estimate the postchange parameters of Hawkes processes and then compute the GLRs. This method requires large memory and computation time to obtain the maximum likelihood estimates at each time instance. An alternative to GLR is the score test (Rao 2005), which does not require computing the maximum likelihood estimation. The score test is well studied and widely applied. In the univariate case, the score test is the most powerful test for small deviation from the null hypothesis. For multistream network data, a strategy is needed to combine the high-dimensional statistic; L. Xie and Xie (2021) proposed a graph scanning statistic by computing the statistics for subgraphs that is helpful to detect and identify local changes. In addition to online change point detection, recent work by D. Wang, Yu, and Willett (2020) developed methods and established theoretical guarantees for offline change point detection for Hawkes processes.

In this article, we present a graph scan score statistic for detecting local changes that happen as a cluster in the network when observing sequential discrete event data that can be modeled using point processes. The change causes an unknown shift in the underlying parameter of the Hawkes process over a subnetwork. We assume that the parameters of the prechange distributions are known because typically abundant "normal" and "in-control" data can be used to estimate the prechange parameters with good precision. We assume that the postchange parameters are unknown because they are typically due to an unexpected anomaly. This motivates us to consider the score statistic, which detects a departure from the "normal" data without having to estimate the postchange parameters. We present the asymptotic distribution of the score statistic, which enables us to develop the self-normalizing scan statistic over predefined candidate scanning clusters. This also leads to an accurate approximation of the instantaneous false alarm probability, the false alarm rate, the average run length, and an efficient procedure to accurately determine the false discovery rate via importance sampling, as validated by numerical examples. The good performance of our procedure compared with the benchmarks is tested with numerical experiments with simulated and real stock exchange data.

The rest of our article is organized as follows. Section 2 provides background knowledge on the multivariate Hawkes process. Section 3 presents the definition of our

problem. Section 4 proposes our detection procedure and includes the analysis of our scan score statistics. Section 5 presents experiments of a simulation study and real-world data application. Section 6 concludes our article.

2. PRELIMINARIES

A multivariate Hawkes process is a self-exciting process over a network. Let M denote the number of nodes in the network and $[M]$ denote $\{1, ..., M\}$. The data are of the form $\{(u_1, t_1), (u_2, t_2), ...\}$, where $u_i \in [M]$ denotes the location of the ith event and $t_i \in (0, \infty)$ denotes the time of the ith event. A multivariate Hawkes process is actually a special case of spatiotemporal counting process (Rathbun 1996). Let $\mathcal{H}_t$ denote the history before time t; that is, the σ-algebras of events before time t; $\{\mathcal{H}_t\}_{t\geq 0}$ is a filtration, an increasing sequence of σ-algebras. Let $N_m(t)$ denote the number of events on the ith node up to time t; that is, a counting process,

$$N_m(t) = \sum_{t_i \leq t} \mathbb{I}(t_i \leq t, u_i = m),$$

where $\mathbb{I}$ denotes the indicator variable. Then, a multivariate Hawkes process can be determined by the following conditional intensity function (Reinhart 2018):

$$\lambda_m(t) = \lim_{s \to 0} \frac{\mathbb{P}\{N_m(t+s) > 0 | \mathcal{H}_t\}}{s}. \tag{2.1}$$

For a multivariate Hawkes process, the conditional intensity function takes the form

$$\lambda_m(t) = \mu_m + \sum_{i \in [M]} \int_0^t g_{i,m}(t-s) N_i(ds). \tag{2.2}$$

Here μ_m is the base intensity and $g_{i,j}(t)$ is the kernel function that characterized the influence of the previous events. Specifically, we assume a commonly used exponential kernel; that is,

$$g_{i,j}(t) = \alpha_{i,j} e^{-\beta t}, \tag{2.3}$$

where $\beta > 0$ is a parameter that controls the decay rate. Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_M)$ and $\mathbf{A} \in \mathbb{R}^{M \times M}$, of which the (i, j)th entry is $\alpha_{i,j} \geq 0$. A multivariate Hawkes process with exponential kernel is parameterized by the base intensity $\boldsymbol{\mu}$, influence matrix $\mathbf{A}$, and decay rate β. Given all events in time window $[0, T]$, the log likelihood function is given by

$$\begin{aligned} \ell_T(\mathbf{A}) = &\sum_{k=1}^{K} \log\left(\mu_{u_k} + \sum_{t_i < t_k} \alpha_{u_i, u_k} e^{-\beta(t_k - t_i)}\right) - \sum_{m=1}^{M} \mu_m T \\ &+ \frac{1}{\beta} \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{u_k, m} \left[e^{-\beta(T - t_k)} - 1\right], \end{aligned} \tag{2.4}$$

where K denotes the number of events before time T. Note that when $\mathbf{A} = \mathbf{0}$, the process becomes a multivariate Poisson process.

3. PROBLEM SETUP

Consider a network with M nodes; we can observe a sequence of events on each node over time. There may exist a change point in time $\tau^* > 0$ if the following applies. Before time τ^*, the events of the network follow a multivariate Hawkes process with parameters $\boldsymbol{\mu}$, $\mathbf{A}_0$, and β. After time τ^*, the events of the network follow a multivariate Hawkes process in which the influence matrix changes from $\mathbf{A}_0$ to $\mathbf{A}_1$. The prechange and postchange multivariate Hawkes processes are assumed to be stationary; that is, $||\mathbf{A}_0|| < 1$ and $||\mathbf{A}_1|| < 1$, where $||\cdot||$ represents the spectral norm. To detect whether a change point τ^* exists in the given data, we consider the following hypothesis test:

$$\begin{aligned} H_0 &: \lambda_m(t) = \mu_m + \textstyle\sum_{t_i \le t} \alpha_{u_i,m,0} e^{-\beta(t-t_i)}, m \in [M], t \ge 0; \\ H_1 &: \lambda_m(t) = \mu_m + \textstyle\sum_{t_i \le \tau^*} \alpha_{u_i,m,0} e^{-\beta(t-t_i)}, m \in [M], 0 \le t \le \tau^*; \\ &\lambda_m(t) = \mu_m + \textstyle\sum_{\tau^* \le t_i \le t} \alpha_{u_i,m,1} e^{-\beta(t-t_i)}, m \in [M], t > \tau^*; \end{aligned} \tag{3.1}$$

where $\lambda_m(t)$ is the true conditional intensity of node m at time t, and $\alpha_{i,j,0}$ and $\alpha_{i,j,1}$ are the (i, j)th entries of $\mathbf{A}_0$ and $\mathbf{A}_1$, respectively. In particular, we refer to $\mathbf{A}_0$ and $\mathbf{A}_1$ as the network parameters because they describe the interactions (influences) between nodes in the network.

4. SCAN SCORE STATISTICS DETECTION PROCEDURE

To perform the sequential hypothesis test (3.1), we proposed a detection procedure based on scan score statistics. A score statistic corresponds to the first-order derivative of the log-likelihood function. In a multivariate Hawkes network, we are interested in the influence between multiple pairs of nodes (i.e., the entries in the influence matrix $\mathbf{A}$). We can derive the score statistics for each pair, which leads to a multidimensional vector of score statistics. Based on that, we use a scanning strategy to compute our test statistics, similar to Chen, Kim, and Xie (2020) and He et al. (2018). Specifically, we divide the network into several clusters, compute the score statistics in each cluster at each time t, and then obtain a detection statistic for each cluster by summing up the standardized score statistics in the corresponding cluster. Finally, we take the maximum over all clusters to form the scan score statistics at time t for the entire network. More details will be discussed in this section.

4.1. Score Statistics

Because the change in hypothesis test (3.1) is caused by the change of influence matrix, we define the following score statistics, given data up to time T, with respect to $\alpha_{p,q}$:

$$S_T^{(p,q)}(\mathbf{A}) \triangleq \frac{\partial \ell_T(\mathbf{A})}{\partial \alpha_{p,q}}. \tag{4.1}$$

Moreover, define $S_T(\mathbf{A})$ as the vector of all elements in $\{S_T^{(p,q)}(\mathbf{A}); p, q \in [M]\}$. According to theorem 1 in Rathbun (1996), we have the following lemma.

Lemma 4.1. *Under the assumptions in Rathbun (1996), assume that the influence matrix of the multivariate Hawkes process is* $\mathbf{A}$. *The score function* $S_T(\mathbf{A})$ *satisfies that*

$$T^{-\frac{1}{2}} S_T(\mathbf{A}) \xrightarrow{D} \mathcal{N}(0, \mathcal{I}(\mathbf{A})),$$

where $\mathcal{I}(\mathbf{A})$ *is the Fisher information.*

4.1.1. Theoretical Characterization of Fisher Information $\mathcal{I}(\mathbf{0})$

When $\mathbf{A} = \mathbf{0}$, it is possible to compute $\mathcal{I}(\mathbf{0})$ as shown in the following theorem. For simplicity, let us define $\mathcal{C}(i,t)$ as the set of events at node i before time t; that is, $\mathcal{C}(i,t) = \{k : t_k < t, u_k = i\}$.

Theorem 4.1. *Assume the conditional intensity function has the form as in equation (2.2) and the kernel function is exponential as in equation (2.3). According to equation (4.1),*

$$\begin{aligned} S_T^{(p,q)}(\mathbf{A}) &= \frac{\partial \ell_T(\mathbf{A})}{\partial \alpha_{p,q}} \\ &= \sum_{k \in \mathcal{C}(q,T)} \frac{\sum_{i \in \mathcal{C}(p,t_k)} e^{-\beta(t_k - t_i)}}{\mu_q + \sum_{t_i < t_k} \alpha_{u_i,q} e^{-\beta(t_k - t_i)}} + \frac{1}{\beta} \sum_{k \in \mathcal{C}(p,T)} \left[e^{-\beta(T - t_k)} - 1 \right]. \end{aligned} \tag{4.2}$$

Moreover, when $\mathbf{A} = \mathbf{0}$, *as* $T \to \infty$, *the nonzero elements in the limit of variance (i.e.,* $\mathcal{I}(\mathbf{A})$ *in Lemma 4.1) are as follows:*

$$\begin{aligned} &\mathrm{Var}\left[T^{-\frac{1}{2}} S_T^{(q,q)}(\mathbf{0}) \right] \to \frac{1}{2\beta} + \frac{\mu_q}{\beta^2}, \\ &\mathrm{Var}\left[T^{-\frac{1}{2}} S_T^{(p,q)}(\mathbf{0}) \right] \to \frac{\mu_p}{\mu_q} \left(\frac{1}{2\beta} + \frac{\mu_p}{\beta^2} \right), \\ &\mathrm{Cov}\left[T^{-\frac{1}{2}} S_T^{(p,q)}(\mathbf{0}), T^{-\frac{1}{2}} S_T^{(p',q)}(\mathbf{0}) \right] \to \frac{\mu_p \mu_{p'}}{\mu_q \beta^2}. \end{aligned} \tag{4.3}$$

4.1.2. Estimation of $\mathcal{I}(\mathbf{A})$

When $\mathbf{A} \neq \mathbf{0}$, it is difficult to compute the variance theoretically. According to Rathbun (1996), we have the following approximation of $\mathcal{I}(\mathbf{A})$.

Theorem 4.2. *With the same assumption as in Theorem 4.1, we have the following estimation of the Fisher information. Let*

$$\hat{\mathcal{I}}_T(\mathbf{A})_{(i,j),(p,q)} = \begin{cases} 0 & \text{if} \quad j \neq q, \\ \sum_{k \in \mathcal{C}(q,T)} \dfrac{\left(\sum_{k \in \mathcal{C}(i,t)} e^{-\beta(t_k - t_i)}\right)\left(\sum_{k \in \mathcal{C}(p,t)} e^{-\beta(t_k - t_i)}\right)}{\left(\mu_q + \sum_{t_i < t_k} \alpha_{u_i,q} e^{-\beta(t_k - t_i)}\right)^2} & \text{if} \quad j = q. \end{cases} \tag{4.4}$$

We have $\frac{1}{T}\hat{\mathcal{I}}_T(\mathbf{A}) \to \mathcal{I}(\mathbf{A})$; *that is,* $\forall i, j, p, q$,

$$\frac{1}{T} \hat{\mathcal{I}}_T(\mathbf{A})_{(i,j),(p,q)} \to \mathcal{I}(\mathbf{A})_{(i,j),(p,q)}, \tag{4.5}$$

where $\mathcal{I}(\mathbf{A})_{(i,j),(p,q)}$ *is the asymptotic (co)variance of* $T^{-1/2} S_T^{(i,j)}(\mathbf{A})$ *and* $T^{-1/2} S_T^{(p,q)}(\mathbf{A})$.

4.2. Scan Score Statistics

To combine all of the score statistics and complete the detection procedure, we compute the scan statistics based on given clusters. A cluster is a directed subgraph, with the set of nodes V_i and the set of edges E_i, $i = 1, 2, ..., L$, where L is the number of clusters. In practice, to reduce the computation cost, we only compute the score statistics given data in a time window of length w and update the statistics every δ time units, where $\delta \leq w$. Specifically, at time t, which is a multiple of δ, for the ith cluster we compute the score statistics on all edges in the cluster at the prechange parameter $\mathbf{A}_0$ with data in $[t - w, t]$ and have a vector of score statistics denoted as $S^{(i)}_{t,w}(\mathbf{A}_0) = (S^{(p,q)}_t(\mathbf{A}_0) - S^{(p,q)}_{t-w}(\mathbf{A}_0))_{(p,q)\in E_i}$. Let R_i and $\mathcal{I}^{(i)}(\mathbf{A}_0)$ denote the dimensions and Fisher information corresponding to edges in E_i, respectively. Then the detection statistic for cluster i at time τ with window length w is

$$\Gamma^{(i)}_{t,w} = (wR_i)^{-1/2}\mathbf{1}^\top \mathcal{I}^{(i)}(\mathbf{A}_0)^{-1/2} S^{(i)}_{t,w}(\mathbf{A}_0) \sim \mathcal{N}(0, 1). \tag{4.6}$$

Note that the scan statistic has a "self-normalizing" property in that the asymptotic distributions are standard normal for multiple candidate clusters, which will facilitate control of the false alarm by choosing a threshold. For example, if we are interested in detecting a shift from a Poisson process, we only need to evaluate $\mathcal{I}(\mathbf{A}_0)$, which only requires the prechange parameter without having to estimate the postchange parameter $\mathbf{A}_1$. The estimation of $\mathbf{A}_1$ can be difficult to perform online given limited postchange observations, because we would like to detect the change quickly.

Then at each time t, we compute the scan score statistics over the candidate clusters:

$$\Gamma_t = \max_{1\leq i\leq L} |\Gamma^{(i)}_{t,w}|.$$

Given a threshold $b > 0$, we stop our procedure and raise an alarm to detect a local change point using the following rule:

$$T_b = \inf\{t : \Gamma_t > b\}. \tag{4.7}$$

In the following, we discuss the false alarm probability at any instant t (Section 4.2.1) and then provide the performance analysis of T_b (Section 4.2.2). Here the choice of the threshold b controls the trade-off between the false alarm rate and average run length versus the detection delay. The proposed method can also be used to localize the change once an alarm is raised, and we briefly discuss the false discovery rate in Appendix A.

4.2.1. Instantaneous False Alarm Probability of Scan Statistics at a Given t

According to Lemma 4.1, the score $T^{-1/2}S_T(\mathbf{A}_0)$ converges in distribution to $\mathcal{N}(0, \mathcal{I}(\mathbf{A}_0))$ as $T \to \infty$. Because the Hawkes process is stationary, $w^{-1/2}(S_{w\tau}(\mathbf{A}_0) - S_{w(\tau-1)}(\mathbf{A}_0))$ also converges to the same distribution as $w \to \infty$. The statistics $\Gamma^{(i)}_{t,w}$ are linear combinations of $w^{-1/2}(S_t(\mathbf{A}_0) - S_{t-w}(\mathbf{A}_0))$, and after scaling the time by w^{-1}, the process $(\Gamma^{(i)}_{w\tau,w})_{\tau\geq 0, 1\leq i\leq L}$ converges pointwise to a Gaussian process as $w \to \infty$, and the covariance can be characterized by

$$\mathrm{Cov}(\Gamma^{(i)}_{w\tau,w},\Gamma^{(j)}_{w(\tau+\epsilon),w})=(1-\epsilon)^{+}\mathrm{Cov}(\Gamma^{(i)}_{1,1},\Gamma^{(j)}_{1,1}),\ \text{for any } 1\le i\le j\le L,\ \text{for any } \epsilon\ge 0, \tag{4.8}$$

where the covariance between $\Gamma^{(i)}_{1,1}$ and $\Gamma^{(j)}_{1,1}$ can be found in closed form when $\mathbf{A}_0=\mathbf{0}$ using Theorem 4.1 and estimated using historical data by Theorem 4.2. At time $t>0$, we want to control the instantaneous false alarm probability

$$\begin{aligned}\mathbb{P}(\Gamma_t>b)&=\mathbb{P}(\max_{1\le i\le L}|\Gamma^{(i)}_{t,w}|\ge b)=\mathbb{P}(\bigcup_{i=1}^{L}|\Gamma^{(i)}_{t,w}|\ge b)\\&=\mathbb{P}(\bigcup_{i=1}^{L}\{\Gamma^{(i)}_{t,w}\ge b\}\bigcup_{i=1}^{L}\{\Gamma^{(i)}_{t,w}\le -b\})\\&=\mathbb{P}(\{\max_{1\le i\le L}\Gamma^{(i)}_{t,w}\ge b\}\cup\{\min_{1\le i\le L}\Gamma^{(i)}_{t,w}\le -b\})\\&\le 2\mathbb{P}(\max_{1\le i\le L}\Gamma^{(i)}_{t,w}\ge b).\end{aligned} \tag{4.9}$$

Let $\mathbf{\Gamma}_{t,w}$ denote the vector of $\Gamma^{(i)}_{t,w}$ s. To control the upper bound of the instantaneous false alarm probability, we compute (4.9) with the technique in Botev, Mandjes, and Ridder (2015):

$$\begin{aligned}\mathbb{P}(\max_{1\le i\le L}\Gamma^{(i)}_{t,w}\ge b)&=\mathbb{P}(\bigcup_{i=1}^{L}\{\Gamma^{(i)}_{t,w}\ge b,\Gamma^{(i)}_{t,w}\ge\Gamma^{(j)}_{t,w},j\ne i\})\\&=\sum_{i=1}^{L}\mathbb{P}(\Gamma^{(i)}_{t,w}\ge b,\Gamma^{(i)}_{t,w}\ge\Gamma^{(j)}_{t,w},j\ne i\})\\&=\sum_{i=1}^{L}\mathbb{P}(\mathbf{B}\mathbf{P}_i\mathbf{\Gamma}_{t,w}\ge\mathbf{b}),\end{aligned} \tag{4.10}$$

where $\mathbf{P}_i$ is the permutation matrix interchanging the first entry and the ith entry, and

$$\mathbf{B}=\begin{bmatrix}1&0&\cdots&\cdots&0\\1&-1&0&\cdots&0\\1&0&-1&\ddots&\vdots\\\vdots&\vdots&\ddots&\ddots&0\\1&0&\cdots&0&-1\end{bmatrix},\quad \mathbf{b}=\begin{pmatrix}b\\0\\\vdots\\0\end{pmatrix}. \tag{4.11}$$

Here $\mathbf{\Gamma}_{t,w}$ follows a Gaussian distribution where the covariance $\mathbf{\Sigma}$ can be computed with (4.8) according to the network topology and the score statistics in the clusters. In Botev, Mandjes, and Ridder (2015), the authors provided an importance sampling algorithm to estimate (4.10). Figure 1 is an example of the cluster structure and the corresponding $\mathbf{\Sigma}$ when $\mathbf{A}_0=\mathbf{0}$.

Remark 4.1. Because our scan statistics are standardized, determining the threshold b with the method in equation (4.10) does not depend on the window length w. However, with a larger w, the Gaussian process approximation would be better. In Table 1, we can see that as the window length increases, the instantaneous false alarm probability is better controlled.

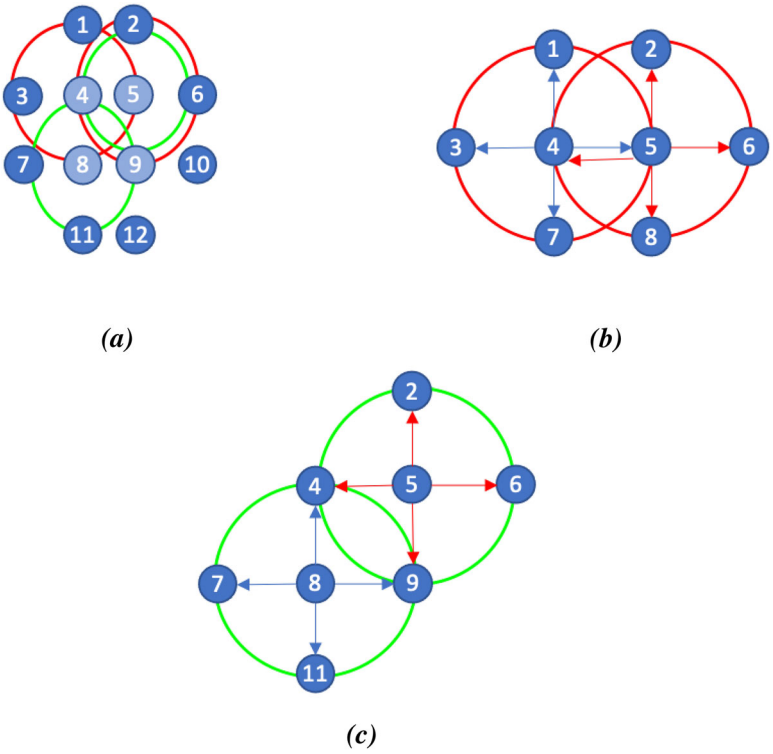

Figure 1. In this example, there are four clusters, and each cluster includes five locations. The four clusters are $(1,3,4,5,8)$, $(2,4,5,6,9)$, $(4,7,8,9,11)$, and $(5,8,9,10,12)$. The light blue nodes are the centers of each cluster and in each cluster we consider the four directions from the center to its neighbors as shown in (b) and (c). $S_{t,w}^{(i)} \sim \mathcal{N}(\mathbf{0}_4, w(1/(2\beta)+\mu/\beta^2)\mathbf{I}_4)$ and $\Gamma_{t,w} \sim \mathcal{N}(0,\Sigma)$. For the case in (b), the Σ_{ij} corresponding to the covariance between $\Gamma_{t,w}^{(i)}$ and $\Gamma_{t,w}^{(j)}$ equals 0. For the case in (c), $\Sigma_{ij} = \sigma^2 \triangleq \mu/(\beta+2\mu)$. Therefore, $\Sigma = ((1,0,0,\sigma^2)^\top, (0,1,\sigma^2,0)^\top, (0,\sigma^2,1,0)^\top, (\sigma^2,0,0,1)^\top)$.

Table 1. Accuracy of approximation of instantaneous false alarm probability through (4.10).

w	b	$2\mathbb{P}(\max_i \Gamma_{t,w}^{(i)} > b)$	$\hat{\mathbb{P}}(\Gamma_t > b)$
50	3	0.01	0.0174
100	3	0.01	0.0146
200	3	0.01	0.0114
50	2.8	0.02	0.0282
100	2.8	0.02	0.0226
200	2.8	0.02	0.0210

4.2.2. Performance Metrics of the Stopping Time T_b

In this part, we provide an upper bound of the false alarm rate (FAR) and average run length (ARL) based on the Gaussian process approximation and the analysis of the instantaneous false alarm probability. Then we discuss the detection delay and the choice of the window length w and update interval δ.

The FAR is the conditional probability that the procedure will stop at the next update, given that there has been no false alarm yet:

$$\mathrm{FAR} = \sup_n (T_b = (n+1)\delta | T_b > n\delta).$$

We have the following result on the FAR.

Theorem 4.3. *As $b \to \infty$, if w/δ is upper bounded by some constant C,*

$$\frac{\text{FAR}}{\mathbb{P}(\Gamma_t > b)} \leq 1 + o(1).$$

The ARL of T_b is $\mathbb{E}[T_b]$, the expected stopping time under H_0. To evaluate the ARL, we are going to show that T_b is an approximately exponential distribution with some parameter λ_0. The analysis is similar to Yakir (2009). Let $f(b) = be^{b^2/2}$, such that $1/f(b)$ is in the same order as the instantaneous false alarm probability. For any $x > 0$ and interval $[0, xf(b)\delta]$, we decompose it into k subintervals with length $m\delta$; that is, $xf(b) = km$. For simplicity, we assume that k and m are integers.

Let indicator X_j denote $\mathbb{I}\{\max_{(j-1)m<n\leq jm}\Gamma_{n\delta} > b\}$, and define $W = \sum_{j=1}^{k} X_j$. Then we have

$$\{W = 0\} = \{T_b > xf(b)\delta\}.$$

To prove that T_b is approximately exponential, it is the same to prove that W is approximately Poisson distributed. We herein apply the result from Arratia, Goldstein, and Gordon (1989). According to theorem I in Arratia, Goldstein, and Gordon (1989), we establish the following theorem.

Theorem 4.4. *Let T_b be the stopping time defined in equation (4.7), X_j be the indicator defined above, and W be the sum of the indicators. As $b \to \infty$, for any $m = 0(f(b))$ which also satisfies $m\frac{\delta}{w} \to \infty$, for any fixed $x \geq 0$,*

$$\lim_{b\to\infty} |\mathbb{P}(T_b > xf(b)\delta) - e^{-\mathbb{E}W}| = 0. \tag{4.12}$$

The theorem above can be used to obtain an approximation of the ARL. According to the construction of W, we have

$$\begin{aligned}\mathbb{E}W &= k\mathbb{P}(X_j = 1)\\ &= xf(b)\mathbb{P}\{\max_{0<n\leq m}\Gamma_{n\delta} > b\}/m\\ &\leq xf(b)\mathbb{P}\{\Gamma_t > b\}.\end{aligned}$$

By Theorem 4.4, $\mathbb{E}_\infty(T_b) \approx \lambda_0^{-1}\delta$ and

$$\lambda_0 \leq \mathbb{P}\{\max_{0<n\leq m}\Gamma_{n\delta} > b\}/m \tag{4.13}$$

$$\leq \mathbb{P}(\Gamma_t > b). \tag{4.14}$$

Therefore, we can use our instantaneous false alarm probability approximation in Section 4.2.1 to approximate (4.13). We can numerically verify that this is a reasonably accurate approximation (see Section 5.1).

Remark 4.2. To evaluate the performance of our scanning statistics, we also need the expected detection delay (EDD). Because we are using a Shewhart chart type of detecting procedure, the EDD varies with the window length. In practice, the window length w can be chosen by considering the smallest change we want to detect on each cluster. Then w will be the smallest window length that has enough power to detect the change

successfully; that is, under the postchange scenario, $|\mathbb{E}(\Gamma_{t,w}^{(i)})| \geq b$ if the change happens on the ith cluster.

Remark 4.3. The performance of our scanning statistic also depends on the update interval δ, where a smaller δ results in both a smaller ARL and a smaller EDD. However, there seems to be little point in choosing δ to be much smaller than w to frequently check for a potential change point while sacrificing computational efficiency because we expect that the change will not be too large and will not be reflected immediately in the detection statistic Γ_t.

5. EXPERIMENTS

5.1. Simulated Result of ARL and EDD

In this experiment, the network is set up as shown in Figure 1. The event in each node follows a Poisson process with $\mu = 1$, and we set $\beta = 1$. The window length is set to be 200, and the statistics are computed for each $\delta = 10$ time units. In Table 2 we show the estimated ARL from simulation with the threshold estimated by (4.13) and (4.14) for $\lambda_0^{-1} \geq 1,000$ and $\lambda_0^{-1} > 2,000$, which corresponds to ARL $\geq \lambda^{-1}\delta$.

To obtain the simulated ARL, we generate events in the time window $[0, 60000]$ and compute the run length when the statistics exceed the corresponding threshold. Note that this approximation will always underestimate the ARL because we can only generate events in a finite time window. We can see that the thresholds computed from (4.13) give us desired results. However, (4.14) tends to overestimate the threshold.

Now, let us compare the EDD of our proposed method with the GLR method in Li et al. (2017). In the experiments for EDD, the distribution under H_0 is set as mentioned above. The thresholds of our methods are set according to the estimate of equation (4.13) with $m = 50$, so that our desired ARLs are 10,000 or 20,000 (see details in Table 2). As for the GLR, we compute the log generalized likelihood ratio with frequency of 0.1 per time unit and window length $w = 200$ with and without the cluster structure. For the GLR with the cluster structure (GLR-C), similar to the proposed method, we compute the statistic on each of the clusters and take the maximum. For the vanilla GLR, we consider a change on the 16 edges in the union of the four clusters. The maximum likelihood estimates for $\mathbf{A}_1$ and $\boldsymbol{\mu}_1$ are computed by the expectation-maximization method. The thresholds of the desired ARLs are estimated with simulation.

We compared the performance of our methods with GLR and GLR-C in different settings, and the results are shown in Table 3. The EDDs are shown in columns 4 to 9 of Table 4. The results show that our proposed method achieves better performance when the change is within cluster and balanced on the edges (Cases 1–3). In Cases 4 and 5,

Table 2. Verification of approximated ARL in (4.13) and (4.14).

	b	Theoretic ARL	Simulated ARL
Results of (4.13), $m = 100$	3.3718	10,000	9,189
Results of (4.13), $m = 50$	3.3859	10,000	9,561
Results of (4.14)	3.6625	10,000	21,773
Results of (4.13), $m = 100$	3.5824	20,000	17,158
Results of (4.13), $m = 50$	3.5867	20,000	17,655
Results of (4.14)	3.8352	20,000	41,701

where the change happens on multiple clusters, and in Case 6, where the change happens on only part of a cluster, the proposed method is comparable to GLR with or without the cluster structure. Table 5 shows the advantage of our method: the computation time of our proposed methods is much less than that for the GLR methods, because our method does not require estimating the postchange distribution parameters. However, in Case 7, if the change happens on a single edge—in other words, the clusters cannot accurately capture the topology of the local change—the performance of the score-based method can be worse than that of the GLR.

Finally, we can verify the exponentiality of the run length by comparing the empirical cumulative distribution function over 500 experiments versus the theoretical one, as shown in Figure 2.

5.2. Real-World Data

In this section, we apply our scan statistics on memetracker data and stock data.

- memetracker data: These data track texts and phrases, which are called "memes," over different websites. These data are used to study information diffusion via social media and blogs. We used three meme data in Li et al. (2017). The first data was "Barack Obama was elected as the 44th president of the United States." We used data from the top 40 news websites, which include Yahoo, CNN, NY Daily, *The Guardian*, etc. We used data from 01 November 2008 to 02 November 2008 as the training data and the data from 3 November 2008 to 5 November 2008 as the test data. Our procedure detected a change at around 7 p.m. on 3 November, which was a few hours before the votes. The second data was "the summer Olympics game in Beijing." We use data from 1 August 2008 to 3 August 2008 as the training data and from 4 August 2008 to 15 August 2008 as the test set.
- Stock price data: These data were downloaded from Yahoo Finance. We collect the closing price and trading volume of stock tickers SPY, QQQ, DIA, EFA, and IWM, which are all index-type stocks and can reflect the situation of the overall stock market. For each ticker, we constructed three types of events. High return: a day with a return over the 90th percentile. Low return: a day with a return below the 10th percentile. High volume: a day with trading volume over the 90th percentile. Therefore, in these data, we have a network with 15 nodes. Such extreme trading events are of interest in the study of finance (Embrechts, Liniger, and Lin 2011). We used data from 4 January 2016 to 31 December 2018 as the training data and from 1 January 2019 to 31 December 2020 as the test data.

Table 3. Setting of different cases in Table 4.

	Changed parameters in postchange distribution
Case 1	$\alpha_{4,1} = \alpha_{4,3} = \alpha_{4,5} = \alpha_{4,8} = 0.2$
Case 2	$\alpha_{4,1} = \alpha_{4,3} = \alpha_{4,5} = \alpha_{4,8} = 0.5$
Case 3	$\alpha_{4,1} = 0.6, \alpha_{4,3} = 0.4, \alpha_{4,5} = \alpha_{4,8} = 0.5$
Case 4	$\alpha_{4,1} = \alpha_{4,3} = \alpha_{9,5} = \alpha_{9,8} = 0.5$
Case 5	$\alpha_{4,5} = \alpha_{4,8} = \alpha_{9,8} = \alpha_{9,5} = 0.5$
Case 6	$\alpha_{4,5} = \alpha_{4,8} = 0.5$
Case 7	$\alpha_{4,5} = 0.5$

Table 4. Comparison of EDD.

Method	Threshold	ARL	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Proposed	3.400	10,000	**104.5**	**44.43**	**46.89**	54.02	45.34	81.92	159.0
GLR-C	7.705	10,000	108.9	48.77	47.77	**51.38**	**39.61**	**68.71**	**97.91**
GLR	11.35	10,000	119.1	52.36	51.73	52.20	41.06	74.23	105.6
Proposed	3.635	20,000	**111.9**	**47.40**	**49.54**	57.82	49.31	89.16	176.9
GLR-C	8.600	20,000	117.0	51.82	50.33	**55.74**	**42.42**	**72.85**	**103.84**
GLR	12.45	20,000	128.2	55.30	55.16	55.85	42.81	79.49	112.5

The bold values mark the best performance under each experimental setup.

Table 5. Comparison of computation time (in seconds) for computing the detection statistic over 50,000 time units.

Method	Proposed	GLR-C	GLR
Duration (s)	3.789	20.47	72.97

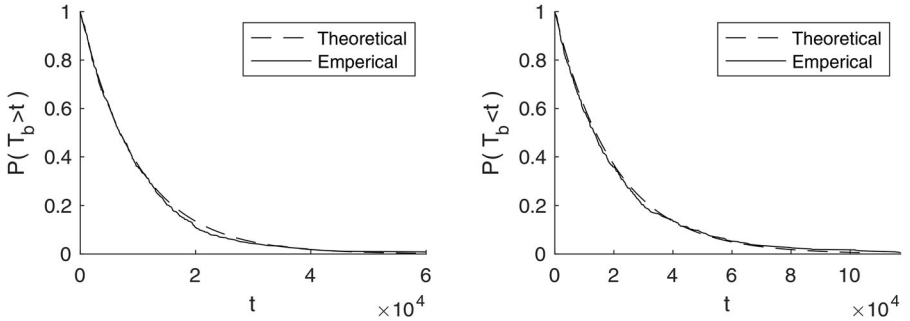

Figure 2. The probability $\mathbb{P}(T_b < t)$ under H_0 at ARL $= 10,000$ and ARL $= 20,000$.

Table 6. Result of real data.

Data	Training set	Test set	No. of cluster	Thresholds	Change points
"Obama"	08.11.01–08.11.02	08.11.03–08.11.05	5	4	3 November, 7 a.m.
"Olympics"	08.08.01–08.08.03	08.08.04–08.08.08	4	4	4 August, 6 p.m.
Stock data	16.01.04–18.12.31	19.01.01–20.12.31	5	4	21 June 2019, 16 August 2019, 4 March 2020

For each data, we applied the Newton method to fit the maximum likelihood estimation of the parameters for the training set. Then we use the fitted parameters to compute the scan statistics on the test set. For memetracker data, we constructed the cluster by applying community detection methods on the fitted $\hat{A}$. For stock data, each cluster included events related to a certain ticker. Details are shown in Table 6. The change point detected from the "Obama" data was around 7 a.m. on 3 November 2008. The result indicates that the public opinion of Barack Obama changed around 1 day before the votes. For the "Olympic" data, our procedure detected a change on 4 August, 3 days before the Olympic games. For the stock data, we detected three time intervals, of which the start dates were 21 June 2019, 16 August 2019, and 4 March 2020. According to the news, the first change point (21 June 2019) was the date that the S&P 500 hit a new record high and the three major stock indexes surged on different scales. The second change point (16 August 2019) was related to the U.S.–China trade war. In August

2019, both the United States and China made multiple announcements about their tariffs. The last change point was 4 March 2020, which was three business days before the first circuit breaker in 2020. There were many change points after the first circuit breaker (9 March 2020), which indicates a long-term change in the stock market caused by the pandemic and trade war. Real data results show that our proposed scan statistics achieve good performance in detecting the real change in different areas such as social media and financial markets (Figure 3).

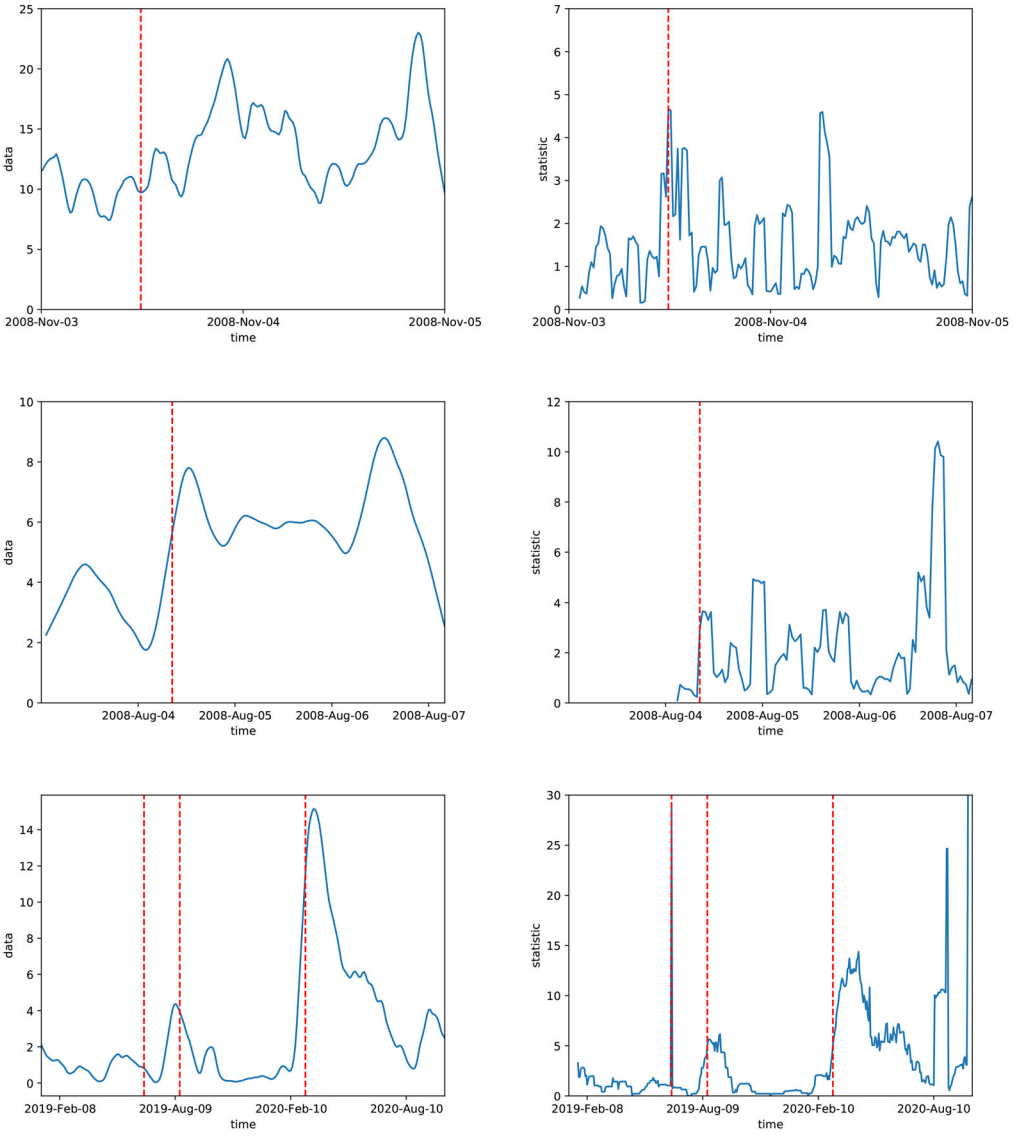

Figure 3. Scan statistics procedure applied on real data. Left column: the blue line is the smoothed frequency of all events in the network. Right column: the blue line is the scan statistics of proposed procedure. First row: data of "Obama". Second row: data of "Olympics". Third row: Stock price data. The dashed red lines indicate the detected change-points.

6. CONCLUSION

In this article, we propose scan score statistics for detecting the change points of network point processes. We use the multivariate Hawkes process to model the sequential event data. Our proposed method is based on score statistics without the requirement to estimate the postchange network parameters, which can be difficult to perform given limited postchange samples because we would like to detect the change quickly. In this sense, our method is more computationally efficient than the conventional GLR method, which is essential in online detection. We derive the asymptotic properties of the scan statistic, which enables us to provide further analysis of the instantaneous false alarm probability, false alarm rate, and average run length and develop a computationally efficient procedure to calibrate the threshold for false alarm control. In experiments, we first use simulated data to verify our theoretical results. We also apply our method using real-world data, and our method shows promising detection performance. Future work includes providing a more detailed discussion of the false discovery rate of localizing the unknown change (some initial results are provided in Appendix A).

APPENDIX A: EXTENSION: FALSE DISCOVERY RATE OF CHANGE LOCALIZATION

This section discusses an extension of the change point detection: false identification after the change detection. After a change has been detected, it is sometimes also of interest to localize it and find the cluster where it happens. This corresponds to multiple hypothesis tests given all of the information up to t. More specifically, at a given t, we check whether $\{|\Gamma_{t,w}^{(i)}| > b\}$ is true or not, for $i = 1, \ldots, L$. Our procedure stops whenever there is at least one discovery; that is, $\exists i$, such that $\{|\Gamma_{t,w}^{(i)}| > b\}$. Let κ denote the number of such discoveries. Among these κ discoveries, there are true discoveries and false discoveries. Let V denote the number of false discoveries. Then the false discovery rate (FDR) is defined as

$$\mathrm{FDR} = \mathbb{E}(V/\kappa; \kappa > 0), \tag{A.1}$$

which is of interest in the study of scanning statistics.

Siegmund, Zhang, and Yakir (2011) provided an estimator for the FDR under the assumptions (i) V is Poisson distributed with expected value ρ and (ii) the number of false discoveries V is independent of the number of true discoveries $\kappa - V$. The estimator is given by

$$\widehat{\mathrm{FDR}} = \rho/(\kappa + 1).$$

In our procedure, assume that L and b are large; with similar proof of Theorem 4.4, we can also show that the first assumption is satisfied; that is,

$$\rho \approx L\mathbb{P}\{|\Gamma_{t,w}^{(i)}| > b\}. \tag{A.2}$$

As for the second assumption, similar to the discussion in Siegmund, Zhang, and Yakir (2011), if each cluster does not largely overlap, most of them are independent. In such a case, the false positives should be an approximately uniform distribution over all clusters. If the true signals do not frequently occur, a false positive is close to a true signal with a very small probability. Therefore, the second assumption is approximately satisfied, too. To control the FDR, we only need to compute the threshold b according to equation (A.2) for a desired ρ. Recall that $\Gamma_{t,w}^{(i)} \sim \mathcal{N}(0, 1)$.

Below we present several simulated examples to demonstrate that the simulated result of the estimation in equation (A.2) provides an accurate approximation of the actual discovery rate. In this experiment, the network contains 20 clusters as shown in Figure A.1. All of the clusters are placed in a line. As in the previous simulated study, the prechange distribution is a Poisson

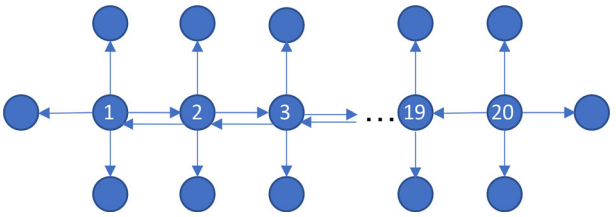

Figure A.1. A network with 20 clusters in a line.

Table 7. Simulated value of false discovery rate, $t = w = 400$, $T_0 = 350$, $T_1 = 50$.

b	$\mathbb{E}(V/(\kappa+1))$	$\mathbb{E}(V)$	$\mathbb{E}(\kappa - V)$	$20\mathbb{P}(\lvert\Gamma^{(i)}_{t,w}\rvert > b)$
1.6	0.530	2.11	0.285	2.19
1.8	0.406	1.35	0.24	1.44
2	0.281	0.8	0.185	0.91
2.2	0.174	0.455	0.15	0.556
2.4	0.113	0.28	0.12	0.328
2.6	0.064	0.155	0.095	0.186
2.8	0.038	0.085	0.075	0.102
3.0	0.026	0.055	0.06	0.054
3.2	0.018	0.035	0.045	0.027
3.4	0.010	0.002	0.035	0.013

Notice that the average number of false discoveries $\mathbb{E}(V)$ is close to the number in the last column, which is the theoretical result.

process with $\mu = 1$. After T_0, the distribution of one cluster changes to a Hawkes process with all the cross-excitation from the center to the neighboring nodes equal to 0.2. We generated the data of total length equal to 400 with $T_0 = 350$ and $T_1 = 50$. All of the clusters are scanned with all the data up to 400. The experiments are repeated 200 times to compute the average number of false discoveries and true discoveries and false discovery rates. Table 7 shows the result. Notice that $\mathbb{E}(V) = \rho$ and is close to the numbers in the last column, which shows a good approximation of equation (A.2).

APPENDIX B: PROOF OF THEOREM 4.1

Proof.

(i) Because under H_0, $S_T^{(q,q)}(0)$ has the same the distribution as the univariate case,

$$\begin{aligned}\mathrm{Var}_{H_0}(T^{-\frac{1}{2}}S_T(0)) &= T^{-1}\mathrm{Var}(S_T(0)) \\ &= T^{-1}\left(\frac{T}{2\beta} + \frac{4\mu T - 1}{4\beta^2} + \frac{e^{-2\beta T}}{4\beta^2} - \frac{3\mu}{2\beta^3} - \frac{\mu e^{-2\beta T}}{2\beta^3} + \frac{2\mu e^{-\beta T}}{\beta^3}\right) \\ &\to \frac{1}{2\beta} + \frac{\mu}{\beta^2} \quad \text{as} \quad T \to \infty.\end{aligned}$$

(ii) To prove the variance of $S_T^{(p,q)}$, we use the fact that $\mathrm{Var}_{H_0}[S_T^{(p,q)}(\mathbf{0})] = -\mathbb{E}_{H_0}\left[\frac{\partial S_T^{(p,q)}(\mathbf{0})}{\partial \alpha_{p,q}}\right]$. Then

$$\begin{aligned}&\mathbb{E}_{H_0}\left[-\frac{\partial S_T^{(p,q)}(\mathbf{0})}{\partial \alpha_{p,q}}\right] \\ &= \mathbb{E}\left[\frac{1}{\mu_q^2}\sum_{k\in\mathcal{C}(q,T)}\left(\sum_{i\in\mathcal{C}(p,t_k)} e^{-\beta(t_k - t_i)}\right)^2\right]\end{aligned} \tag{B.1}$$

$$= \mathbb{E}\left[\frac{1}{\mu_q^2}\mathbb{E}\left[\sum_{k\in\mathcal{C}(q,T)}\left(\sum_{i\in\mathcal{C}(p,t_k)} e^{-\beta(t_k-t_i)}\right)^2 | N_q, N_p\right]\right] \tag{B.2}$$

$$= \mathbb{E}\left[\frac{N_q}{\mu_q^2}\mathbb{E}\left[\left(\sum_{i=1}^{N_p} Z_i(t_k)\right)^2 | N_p, t_k\right]\right] \tag{B.3}$$

$$= \mathbb{E}\left[\frac{N_q}{\mu_q^2}\mathbb{E}\left[\sum_{i=1}^{N_p} Z_i^2(t_k) + \sum_{i\neq j}^{N_p} Z_i(t)Z_j(t) | N_p, t_k\right]\right]$$

$$= \mathbb{E}\left[\frac{N_q}{\mu_q^2}\left(N_p\mathbb{E}\left[Z_i^2(t_k)\right)|t_k\right] + N_p(N_p-1)\mathbb{E}_{i\neq j}\left[Z_i(t_k)Z_j(t_k)|t_k\right]\right)\right] \tag{B.4}$$

$$= \mathbb{E}\left[\frac{N_q}{\mu_q^2}\left(\frac{N_p}{2\beta T}(1-e^{-2\beta t_k}) + \frac{N_p(N_p-1)}{\beta^2T^2}(1-e^{-\beta t_k})^2\right)\right]$$

$$= \frac{T}{\mu_q}\left(\frac{1}{2\beta} + \frac{\mu_q}{\beta^2}\right) + o(T), \tag{B.5}$$

where N_q and N_p are the number of events in $[0, T]$ on nodes q and p, respectively. In equation (B.2), we use the fact that for Poisson processes, the arrival times follow an independent and identically distributed uniform distribution when it is conditional on the number of arrivals. With this fact, in equation (B.3), we define

$$Z_i(t) = \begin{cases} 0 & \text{if } t_i \geq t, \\ e^{-\beta(t-t_i)} & \text{if } t_i < t. \end{cases}$$

Because $t_i \overset{i.i.d}{\sim} \text{unif}[0, T]$,

$$\mathbb{E}Z_i(t) = \frac{1}{T}\int_0^t e^{-\beta(t-u)}du = \frac{1}{\beta T}(1-e^{-\beta t})$$

$$\mathbb{E}Z_i^2(t) = \frac{1}{T}\int_0^t e^{-2\beta(t-u)}du = \frac{1}{2\beta T}(1-e^{-2\beta t}),$$

which proves equation (B.4). Because N_p and N_q follow a Poisson distribution with mean $T\mu_p$ and $T\mu_q$, respectively, equation (B.5) is proved.

(iii) Following a similar technique as in (ii), we can prove

$$\text{Cov}_{H_0}\left[T^{-\frac{1}{2}}S_T^{(p,q)}(\mathbf{0}), T^{-\frac{1}{2}}S_T^{(p',q)}(\mathbf{0})\right] \to \frac{\mu_p\mu_{p'}}{\mu_q\beta^2}.$$

□

APPENDIX C: PROOF OF THEOREM 4.2

Proof. Following the definition in Rathbun (1996), let us define the kernel function,

$$g(\mathbf{s}_1, \mathbf{s}_2, t) = \mathbf{s}_2^\top A\mathbf{s}_1 e^{-\beta t},$$

where $\mathbf{s}_i \in \mathbb{R}^M$ for $i = 1, 2$. Then we can define the conditional intensity function:

$$\Lambda(\mathbf{s},t) = \mu(\mathbf{s}) + \int_0^t \int_X g(\mathbf{s},\mathbf{u},t-r)N(d\mathbf{u}\times dr)$$
$$= \mu(\mathbf{s}) + \sum_{t_i<t} \mathbf{u}_i^\top A\mathbf{s}\cdot e^{-\beta(t-t_i)},$$

where $\mathbf{u} = e_m$ if $u_i = m$, and e_m is the vector where the mth entry is 1 and other entries are 0. Further, define a measure with the delta function

$$\nu(x) = \sum_{i=1}^{M} \delta_{e_i}(x)$$
$$\delta_{e_i}(x) = \begin{cases} 1 & \text{if} \quad x = e_i \\ 0 & \text{o.w.} \end{cases}$$

We can write the likelihood function as follows:

$$\ell_T(A) = \int_0^T \int_X \log\Lambda(\mathbf{s},t;A)N(d\mathbf{s}\times dt) - \int_0^T \int_X \Lambda(\mathbf{s},t;A)\nu(d\mathbf{s})dt.$$

We can easily check that this defines the same multivariate Hawkes process in equations (2.2)–(2.4). Define the matrix $\Delta \in \mathbb{R}^{M^2\times M^2}$ as

$$\Delta_{(i,j),(p,q)} \triangleq \frac{\dot\Lambda_{i,j}\dot\Lambda_{p,q}}{\Lambda} \quad \forall i,j,p,q \in [M],$$

where $\dot\Lambda_{i,j}$ is the partial derivative of Λ with respect to $\alpha_{i,j}$. Therefore, by the result in equation (4.7) of Rathbun (1996), we have

$$\frac{1}{T}\sum_{k=1}^{K} \frac{\Delta(\mathbf{u}_k,t_k)}{\Lambda(\mathbf{u}_k,t_k)} \to \mathcal{I}(A).$$

By direct computation, we have the result of equation (4.4). □

APPENDIX D: PROOF OF THEOREM 4.3

Proof. For every $n \geq w/\delta - 1$, let $p_n = \mathbb{P}(T_b = (n+1)\delta | T_b > n\delta)$. Then FAR $= \sup_n p_n$. For the smallest n, clearly there is $p_n = \mathbb{P}(\Gamma_{n\delta} > b)$ equal to the instantaneous false alarm probability. Also, for every n, there is

$$p_n = \frac{\mathbb{P}(T_b = (n+1)\delta)}{\mathbb{P}(T_b > n\delta)} \leq \frac{\mathbb{P}(T_b > (n+1)\delta - w, \Gamma_{(n+1)\delta} > b)}{\mathbb{P}(T_b > n\delta)}.$$

Because $\{T_b > (n+1)\delta - w\}$ and $\{\Gamma_{(n+1)\delta > b}\}$ are independent, we have

$$p_n \leq \frac{\mathbb{P}(T_b > (n+1)\delta - w)}{\mathbb{P}(T_b > n\delta)}\mathbb{P}(\Gamma_{(n+1)\delta} > b) = \frac{\mathbb{P}(\Gamma_t > b)}{\prod_{k=n+1-\lceil w/\delta\rceil}^{n-1}(1-p_k)}.$$

As $b \to \infty$, the instantaneous false alarm probability $\mathbb{P}(\Gamma_t > b)$ goes to 0, and for large enough b, there exists some $p^* \geq \mathbb{P}(\Gamma_t > b)$, $p^* = \mathbb{P}(\Gamma_t > b)(1+o(1))$, such that $\mathbb{P}(\Gamma_t > b)(1-p^*)^{1-C} = p^*$. Then by induction we can see that for every n,

$$p_n \leq \frac{\mathbb{P}(\Gamma_t > b)}{\prod_{k=n+1-\lceil w/\delta\rceil}^{n-1}(1-p_k)} \leq \frac{\mathbb{P}(\Gamma_t > b)}{(1-p^*)^C} = p^*.$$

□

APPENDIX E: PROOF OF THEOREM 4.4

Proof. According to theorem I in Arratia, Goldstein, and Gordon (1989), let us define the "neighbor of dependence" for index j, $J(j) = \{(j-1), j, j+1\}$, with simple modification for $j=1$ and $j=k$, for $m > w$, where X_j and X_i are independent for $i \notin J(j)$. Therefore, the dependence of elements not in the neighbor vanished; that is, b_3 and b_3' equal 0.

$$\begin{aligned} b_1 &= \sum_{j=1}^{k} \sum_{i \in J(j)} \mathbb{P}(X_j = 1)\mathbb{P}(X_i = 1) \\ &\leq 3k\mathbb{P}(X_1 = 1)^2, \\ b_2 &= \sum_{j=1}^{k} \sum_{i \in J(j)\setminus j} \mathbb{P}(X_j = 1, X_i = 1) \\ &\leq 2k\mathbb{P}(X_1 = 1, X_2 = 1). \end{aligned} \tag{E.1}$$

Note that the event $\{X_1 = 1\}$ is the union of $\{\max_{0<n\leq m-w/\delta} \Gamma_{n\delta} \geq b\}$ and $\{\max_{m-w/\delta<n\leq m} \Gamma_{n\delta} \geq b\}$, and the former is independent of X_2. The same decomposition can be done to X_2. Then b_2 can be upper bounded by

$$\begin{aligned} b_2 &\leq 2k\left(\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) + \mathbb{P}\left(\max_{m-w/\delta<n\leq m} \Gamma_{n\delta} \geq b, \max_{m<n\leq m+w/\delta} \Gamma_{n\delta} \geq b\right)\right) \\ &\leq 2k\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) + 2k\mathbb{P}\left(\max_{m-w/\delta<n\leq m} \Gamma_{n\delta} \geq b\right). \end{aligned} \tag{E.2}$$

With the inequality of the tail probability of normal distribution in Feller (1957),

$$\mathbb{P}(X_1 = 1) = \mathbb{P}\left\{\max_{\substack{0<n\leq m,\\ 1\leq i\leq L}} |\Gamma^{(i)}_{n\delta,w}| > b\right\} \leq 2mL\bar{\Phi}(b) \leq \frac{2mL}{b} e^{-\frac{b^2}{2}},$$

where $\bar{\Phi}$ is the tail probability of a standard normal random variable. With the same computation, we can show

$$\mathbb{P}\left(\max_{m-w/\delta<n\leq m} \Gamma_{n\delta} \geq b\right) \leq \frac{2wL/\delta}{b} e^{-\frac{b^2}{2}}.$$

Therefore, with theorem I in Arratia, Goldstein, and Gordon (1989),

$$|\mathbb{P}(T_b > xf(b)\delta) - e^{-\mathbb{E}W}| \tag{E.3}$$

$$= |\mathbb{P}(W = 0) - e^{-\mathbb{E}W}| \tag{E.4}$$

$$< b_1 + b_2 \tag{E.5}$$

$$\leq \frac{12km^2L^2}{b^2e^{b^2}} + \frac{8km^2L^2}{b^2e^{b^2}} + \frac{4kwL/\delta}{be^{b^2/2}} \tag{E.6}$$

$$= \frac{12xmL^2}{be^{b^2/2}} + \frac{8xmL^2}{be^{b^2/2}} + \frac{4xwL/\delta}{m}, \tag{E.7}$$

which becomes small when $b \to \infty$. □

ACKNOWLEDGMENT

The authors thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this article.

DISCLOSURE

The authors have no conflicts of interest to report.

FUNDING

This work is partially supported by an NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210.

ORCID

Yao Xie http://orcid.org/0000-0001-6777-2951

REFERENCES

Arratia, R., L. Goldstein, and L. Gordon. 1989. "Two Moments Suffice for Poisson Approximations: The Chen-Stein Method." *The Annals of Probability* 17 (1):9–25. doi:10.1214/aop/1176991491

Botev, Z. I., M. Mandjes, and A. Ridder. 2015. "Tail Distribution of the Maximum of Correlated Gaussian Random Variables." 2015 Winter Simulation Conference (WSC), Huntington Beach, CA, USA, 633–42. doi:10.1109/WSC.2015.7408202

Chen, J., S. H. Kim, and Y. Xie. 2020. "S 3 T: A Score Statistic for Spatiotemporal Change Point Detection." *Sequential Analysis* 39 (4):563–92. doi:10.1080/07474946.2020.1826796

Christakis, N. A., and J. H. Fowler. 2010. "Social Network Sensors for Early Detection of Contagious Outbreaks." *PLoS One* 5 (9):e12948. doi:10.1371/journal.pone.0012948

Embrechts, P., T. Liniger, and L. Lin. 2011. "Multivariate Hawkes Processes: An Application to Financial Data." *Journal of Applied Probability* 48 (A):367–78. doi:10.1239/jap/1318940477

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications*. Hoboken, NJ: John Wiley & Sons.

Hawkes, A. G. 1971. "Spectra of Some Self-Exciting and Mutually Exciting Point Processes." *Biometrika* 58 (1):83–90. doi:10.1093/biomet/58.1.83

Hawkes, A. G. 2018. "Hawkes Processes and Their Applications to Finance: A Review." *Quantitative Finance* 18 (2):193–8. doi:10.1080/14697688.2017.1403131

He, X., Y. Xie, S. M. Wu, and F. C. Lin. 2018. "Sequential Graph Scanning Statistic for Change-Point Detection." 2018 52nd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 1317–21. doi:10.1109/ACSSC.2018.8645505

Li, S., Y. Xie, M. Farajtabar, A. Verma, and L. Song. 2017. "Detecting Changes in Dynamic Events over Networks." *IEEE Transactions on Signal and Information Processing over Networks* 3 (2):346–59. doi:10.1109/TSIPN.2017.2696264

Moustakides, G. V. 2008. "Sequential Change Detection Revisited." *The Annals of Statistics* 36 (2):787–807. doi:10.1214/009053607000000938

Ogata, Y. 1998. "Space-Time Point-Process Models for Earthquake Occurrences." *Annals of the Institute of Statistical Mathematics* 50 (2):379–402. doi:10.1023/A:1003403601725

Poor, H. V., and O. Hadjiliadis. 2008. *Quickest Detection*. Cambridge, England: Cambridge University Press.

Rao, C. R. 2005. "Score Test: Historical Review and Recent Developments." In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, edited by N. Balakrishnan, H. N. Nagaraja and N. Kannan, 3–20. New York, NY: Springer.

Rathbun, S. L. 1996. "Asymptotic Properties of the Maximum Likelihood Estimator for Spatio-Temporal Point Processes." *Journal of Statistical Planning and Inference* 51 (1):55–74. doi:10.1016/0378-3758(95)00070-4

Reinhart, A. 2018. "A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications." *Statistical Science* 33 (3):299–318. doi:10.1214/17-STS629

Rizoiu, M. A., Y. Lee, S. Mishra, and L. Xie. 2017. "Hawkes Processes for Events in Social Media." *Frontiers of Multimedia Research*, 191–218.

Rizoiu, M. A., S. Mishra, Q. Kong, M. Carman, and L. Xie. 2018. "SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Populations." In Proceedings of 2018 World Wide Web Conference, Lyon, France, 419–28.

Siegmund, D. O., N. R. Zhang, and B. Yakir. 2011. "False Discovery Rate for Scanning Statistics." *Biometrika* 98 (4):979–85. doi:10.1093/biomet/asr057

Tartakovsky, A. G. 2019. *Sequential Change Detection and Hypothesis Testing: General Non-iid Stochastic Models and Asymptotically Optimal Rules*. Boca Raton, FL: Chapman and Hall/CRC.

Veeravalli, V. V., and T. Banerjee. 2014. "Quickest Change Detection." In *Academic Press Library in Signal Processing*, Vol. 3, 209–55. Amsterdam, Netherland: Elsevier.

Wang, D., Y. Yu, and R. Willett. 2020. "Detecting Abrupt Changes in High-Dimensional Self-Exciting Poisson Processes." *arXiv* preprint arXiv: 2006.03572.

Wang, H., L. Xie, Y. Xie, A. Cuozzo, and S. Mak. 2022. "Sequential Change-Point Detection for Mutually Exciting Point Processes over Networks." Accepted, *Technometrics* and *arXiv* preprint arXiv: 2102.05724.

Xie, L., and Y. Xie. 2021. "Optimality of Graph Scanning Statistic for Online Community Detection." 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Victoria, Australia, 386–90. doi:10.1109/ISIT45174.2021.9518257

Xie, L., S. Zou, Y. Xie, and V. Veeravalli. 2021. "Sequential (Quickest) Change Detection: Classical Results and New Directions." *IEEE Journal on Selected Areas in Information Theory* 2 (2):494–514. doi:10.1109/JSAIT.2021.3072962

Xie, Y., and D. Siegmund. 2013. "Sequential Multi-Sensor Change-Point Detection." *The Annals of Statistics* 41 (2):670–92. doi:10.1214/13-AOS1094

Yakir, B. 2009. "Multi-Channel Change-Point Detection Statistic with Applications in DNA Copy-Number Variation and Sequential Monitoring." In Proceedings of Second International Workshop in Sequential Methodologies, UTT, Troyes, France, 15–7.