Detecting weak changes in dynamic events over networks

Shuang Li * Yao Xie † Mehrdad Farajtabar ‡ Le Song §

Abstract

Large volume of event data are becoming increasingly available in a wide variety of applications, such as social network analysis, Internet traffic monitoring and healthcare analytics. Event data are observed irregularly in continuous time, and the precise time interval between two events carries a great deal of information about the dynamics of the underlying systems. How to detect changes in these systems as quickly as possible based on such event data?

In this paper, we present a novel online detection algorithm for high dimensional event data over networks. Our method is based on a likelihood ratio test for point processes, and achieve weak signal detection by aggregating local statistics over time and networks. We also design an online algorithm for efficiently updating the statistics using an EM-like algorithm, and derive highly accurate theoretical characterization of the false-alarm-rate. We demonstrate the good performance of our algorithm via numerical examples and real-world twitter and memetracker datasets.

Index terms— Change-point Detection, Point Processes, Online Algorithm

1 Introduction

Networks have become a convenient tool for people to efficiently disseminate, exchange and search for information. Networks also provide a rapid way to propagate virus, rumors and diseases, which greatly threat users’ privacy and safety. Recent attacks on very popular web sites such as Yahoo and eBay [12], leading to a disruption of services to users, have triggered an increasing interest in network anomaly detection. Once such an anomaly appears, it will be harder to control as time goes by if not detect and taken actions in time. In the positive side, surge of hot topics and breaking news can provide business opportunities. Missing out such opportunities will result in losing business competitive edge. Therefore, early detection of anomalies, epidemic outbreaks, hot topics, or new trends among streams of data from networked entities is a very important task and has been attracting significant interests [2][12,13].

*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. Email: sli370@gatech.edu
†H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. Email: yao.xie@isye.gatech.edu
‡College of Computing, Georgia Institute of Technology. Email: mehrdad@gatech.edu
§College of Computing, Georgia Institute of Technology. Email: lsong@cc.gatech.edu
Figure 1: Asynchronously and interdependently generated high dimensional event data are fundamentally different from i.i.d. and time-series data. First, observations for each dimension can be collected at different time points. Second, there can be temporal dependence as well as cross-dimensional dependence. In contrast, the dimensions of i.i.d. and time-series data are sampled at the same time point, and in the figure, different marks indicate potentially different values or features of an observation.

Nowadays, data from networked entities are dynamic in nature, and many of them are discrete events observed irregularly in continuous time \cite{17,22}. The precise time interval between two events carries a great deal of information about the dynamics of the underlying systems. These characteristics make such event data fundamentally different from independently and identically distributed (i.i.d.) data, and time-series data where time and space is treated as an index rather than random variables (see Figure 1 for further illustrations of the distinctive nature of event data vs. i.i.d. and time series data.) Obviously, i.i.d. assumption can not capture temporal dependency between data points, while time-series models require us to discretize the time axis and aggregate the observed events into bins. It is not clear how one can choose the size of the bin and how to best deal with the case when there is no event within a bin. Given such distinctive nature of the event data, then how to detect changes and take into account the temporal dependency between these events?

Besides the distinctive temporal aspect, there are three additional challenges for change point detection for event data: (i) how to detect weak signals over networks, (ii) how to update the statistics efficiently online and (iii) how to provide theoretical characterization of the false-alarm-rate for the statistics.

For the first challenge, many existing approaches usually use random or ad-hoc aggregations which may not pool data efficiently and hence loses statistical power to detect weak signals. Take social network as an example: each user in the social network would has their own behavior patterns which would vary individually; at the same time every user in the network would be inevitably influenced by their neighbors and would alternating her behavior correspondingly. Since the anomalies or new topics are initially weak and hidden in the social networks, smart aggregation methods are needed to pick up weak signals \cite{16}.

For the second challenge, many existing detection methods based on likelihood ratio statistics do not take into account computational complexity and are not scalable to large networks. Temporal events can arrive at the Internet servers or social platforms in very high volume and velocity. For instance, every day on average around 500 million tweets are tweeted on Twitter\cite{1}. Thus there is a great need for developing efficient algorithms for update the detection statistics online.

For the third challenge, it is usually very hard to control false-alarm for change-point

\footnote{http://www.internetlivestats.com/twitter-statistics/}
detection statistics over a large network. It has been observed that when applied to real
network data, traditional detection approaches usually have a high false alarms \[12\]. This
would lead to a huge waste of resources since every time an anomaly is declared, subsequent
diagnoses are needed. Lacking accurate theoretical characterization of false-alarms, existing
approaches usually have to incur expensive Monte Carlo simulations to determine the false-
alarms and is prohibitive for large networks.

**Our contributions.** In this paper, we present a novel online detection algorithm tailored
to high-dimensional intertwined event data streams over networks. We will formulate the
change-point detection problem using the mathematical framework of point processes where
before the change the event stream follows one point processes (e.g., Poisson process or
Hawkes process), and after the change the event stream becomes a different point process.
Our goal is to detect such a change *as quickly as possible after it occurs*. In our setting,
we assume the network topology is known and fixed, and the network topology is taken
into account in the stochastic models of the event streams. In particular, our contributions
include the following:

(i) We present a new likelihood ratio based approach for detecting changes for point
processes over networks. Our model explicitly captures the information diffusion both over
networks and over time, and allows us to aggregate information for weak signal detection.

(ii) Our method allows for a general transitions such as Poisson to Hawkes processes, or
Hawkes to Hawkes process. Previous state-of-art usually assumes relatively simple Poisson
model \[29\].

(iii) We present an efficient expectation-maximization like algorithm for updating the
likelihood-ratio detection statistic online.

(iv) We also present sharp theoretical approximation to the false-alarm-rate of the de-
tection algorithm, via the recently developed change-of-measure approach to handle highly
correlated statistics. Our theoretical approximation can be used to determine the threshold
in the algorithm accurately. Finally, we show that the analytical approximation for the false-
alarm-rate can be evaluated directly for simple network topologies such as the star networks
or the chain networks; for general topology network, our false-alarm-rate expression provides
a viable interface for simulation approaches. Using both synthetic and real world datasets,
we show that our online detection algorithm is very sensitive to true changes, and the the-
oretical false-alarm-rates are very accurate compared to the experimental results. Detailed
proofs of theorems are delegate to Appendix.

2 Related work

Recently, there has been a surge of interests in using multidimensional point processes for
modeling dynamic event data over networks. However, many of these works focus on model-
ing networks of point processes and estimating network structures and parameters offline or
even online, including modeling bursty dynamics \[17\]; shaping social activity by incentiviza-
tion \[5\]; learning information diffusion networks \[22\] and source identification from cascades
of events \[6\]; learning mutually exciting processes for viral diffusion \[28\]; learning triggering
kernels for multi-dimensional Hawkes processes \[30\]; in networks where each dimension is a
Poisson process \[20\]; learning latent network structure for general counting processes \[15\];
tracking parameters of dynamic point process networks \[9\]; estimating point-process models of social network interactions \[31\]; and estimating point process models for the co-evolution of network structure an information diffusion \[7\].

Detection of changes in point processes has so far mostly focused on a single event stream rather than the multidimensional case with networks, including change in the intensity of a Poisson process \[24, 29, 11\]; detecting change using the self-exciting Hawkes processes include trend detection in social networks \[19\]; inferring leadership in e-mail networks \[8\]; detecting for Poisson processes using a score statistic \[25\]; detecting changes in Gaussian processes via \[23\]. In neural spike chain change-detection \[21\], data is preprocessed by be counting the number of events over discrete time intervals. Such an ad-hoc approach can not handle the continuous data stream directly, which may distort the distribution of the data and can not detect quickly.

3 Temporal Point Processes

A temporal point process is a random process whose realization consists of a list of discrete events localized in time, \( \{t_i\} \), with \( t_i \in \mathbb{R}^+ \) and \( i \in \mathbb{Z}^+ \). Let \( N_t \) represent the number of events before time \( t \). Let the history \( \mathcal{H}(t) \) be the list of times of events \( \{t_1, t_2, \ldots, t_n\} \) up to but not including time \( t \). Then the number of observed events in a small time window \( dt \) between \( [t, t + dt) \) is \( dN_t = \sum_{t_i \in \mathcal{H}(t)} \delta(t - t_i) \), where \( \delta(t) \) is a Dirac delta function. The total number of observed events till time \( t \) is \( N_t = \int_0^t dN_s \).

For Poisson and inhomogeneous Poisson processes, the intensity is deterministic, which is simply \( \lambda_t = \mu_t \). Hawkes processes and their multidimensional extensions are designed to capture the self and mutual excitation nature of the events in social networks. They have found many applications as we discussed in the related work section. More specifically, one-dimensional Hawkes processes have intensity function as

\[
\lambda_t = \mu_t + \alpha \int_{-\infty}^{t} \varphi(t - \tau) dN_\tau, \tag{1}
\]

where \( \mu_t \) is the base intensity (deterministic) and the influence parameter \( \alpha \in (0, 1) \) (due to the requirement of stationary condition) and the kernel function \( \varphi(t) : \int \varphi(t) dt = 1 \), together characterize how the counting process \( N_t \) influences the intensity. Therefore, the intensity for Hawkes processes is stochastic, which would depend on history. Commonly used kernel function is the exponential kernel \( \varphi(t) = \beta e^{-\beta t} \) which we will use through the paper.

The definition of one-dimensional Hawkes processes can be easily extended to high dimensional cases. We introduce a multivariate counting process \( (N^1_t, N^2_t, \ldots, N^d_t) \), \( t \geq 0 \), with each component \( N^i_t \) recording the number of events of the \( i \)-th component (node) of the network during \( [0, t] \). The intensity function is \( \lambda^i_t = \mu^i_t + \sum_{j=1}^{d} \alpha_{ij} \varphi(t - \tau) dN^j_\tau \), where \( \alpha_{ij}, j, i \in \{1, \ldots, d\} \) indicates the strength of influence of \( N^j \) on \( N^i \) by affecting its intensity process \( \lambda^i \). If \( \alpha_{ij} = 0 \), then it means that \( N^j \) is not influencing \( N^i \).
4 Algorithm

We will focus on two scenarios: the reference data is a family of \textit{i.i.d.} Poisson processes and we want to detect the emergence of self- or mutual-excitation; the reference data are correlated one-dimensional or high-dimensional Hawkes processes and our objective is to detect the changes in self- or mutual-excitation. These two settings have wide applications in real world, and can be generalized to other combination of point processes before and after change-points.

We construct likelihood ratio test to detect the changes from the reference model to the alternative model. Likelihood ratio test is known to be the most powerful test if no unknown parameters in reference and alternative models, which is justified by the Neyman-Pearson lemma \cite{18}.

4.1 Changes from Poisson process to Hawkes process

To detect the changes from Poisson processes to Hawkes processes, we construct the hypothesis test as:

\begin{align}
H_0 : & \quad \lambda_s = \mu, \quad 0 < s < t; \\
H_1 : & \quad \lambda_s = \mu, \quad 0 < s < \tau; \\
& \quad \lambda_s^* = \mu + \alpha \int_{\tau}^{s} \varphi(s-v) dN_v, \quad \tau \leq s < t. \quad (2)
\end{align}

Then we can write out the log-likelihood function based on the data stored in between the time interval \((0, t)\). Under the null, the log-likelihood function \(L_{H_0}\) is

\[
\int_{0}^{t} \log \lambda_s dN_s - \int_{0}^{t} \lambda_s ds. \quad (3)
\]

Under the alternative, the log-likelihood function \(L_{H_1}\) is

\[
\int_{0}^{\tau} \log \lambda_s dN_s + \int_{\tau}^{t} \log \lambda_s^* dN_s - \int_{0}^{\tau} \lambda_s ds - \int_{\tau}^{t} \lambda_s^* ds. \quad (4)
\]

Hence, the log-likelihood ratio is:

\[
\ell = L_{H_1} - L_{H_0} = \int_{\tau}^{t} \log \left( \frac{\lambda_s^*}{\lambda_s} \right) dN_s - \int_{\tau}^{t} (\lambda_s^* - \lambda_s) ds. \quad (5)
\]

Note that \((5)\) is the general log-likelihood ratio for any two point processes with intensity \(\lambda_s\) and \(\lambda_s^*\) respectively. In particular, if we plug in the intensity defined by \((2)\), choose \(\varphi(t) = \beta e^{-\beta t}\), where \(\beta > 0\), and define \(L = t - \tau\), we can compute the log-likelihood ratio as

\[
\ell = \sum_{i=1}^{n} \log \left[ \mu + \alpha \sum_{\tau < t_i < t} \beta e^{-\beta(t_i - t)} \right] - n \log \mu - \alpha \sum_{i=1}^{n} \left[ 1 - e^{-\beta(t - t_i)} \right], \quad (6)
\]

where \(n\) is the number of events within time interval \((\tau, t)\).

Suppose intensity \(\mu\) can be estimated from the reference data and \(\beta\) is given as a priori.
From (6), we see the online log-likelihood ratio only depends on the events in interval \((\tau, t)\) and \(\alpha\). We therefore denote \(\ell\) as \(\ell_{t, \tau, \alpha}\), meaning it is parameterized by \(t, \tau\) and \(\alpha\). Furthermore, we can get the maximum likelihood estimator (MLE) of \(\alpha\) from events in \((\tau, t)\). Even though there does not exist a closed-form estimator for \(\alpha\), we can easily estimate \(\alpha\) via EM like algorithm, which will be discussed in section 4.3.

We will declare that there is a change-point whenever
\[
\sup_{t, \tau, \alpha} \ell_{t, \tau, \alpha} > x, 
\]  
where \(x\) is the pre-determined threshold. Given a sequence of \(n\) observations, define \(\hat{\alpha} = \arg\sup_\alpha \ell_{t, \tau, \alpha}\) is the MLE for \(\alpha\). We can also write (7) as \(\sup_{t, \tau} \ell_{t, \tau, \hat{\alpha}} > x\).

Follow the same framework, we can establish the likelihood ratio for high-dimensional case and when the null distribution is a Poisson process. We construct the likelihood ratio test to detect the events changing from a family of \(i.i.d\). Poisson processes to intertwined multivariate Hawkes processes. The network structure is automatically embedded in the constructed statistic and the information among all the nodes are aggregated by the single statistic.

Written in matrix form, we consider the following hypothesis test:
\[
H_0: \quad \lambda_s = \mu, \quad 0 < s < t; \\
H_1: \quad \lambda_s = \mu, \quad 0 < s < \tau; \\
\lambda^*_s = \mu + A \int^s_\tau \varphi(s-v) dN_v, \quad \tau \leq s < t, 
\]  
where \(dN_v = [dN^1_v, dN^2_v, \ldots, dN^d_v]^{\top}\), \(\mu = [\mu^1, \mu^2, \ldots, \mu^d]^{\top}\), \(A = [\alpha_{ij}]_{1 \leq i, j \leq d}\), and \(\alpha_{ij}\) represents the influence from component \(j\) to \(i\). Such a hypothesis test can be used to detect the emergence of a community that level of self- and mutual-excitation among network increase after the change-point.

Similarly, we define \(e = [1, 1, \ldots, 1]^{\top}\) and obtain the log-likelihood ratio on the basis of (8) as
\[
\ell_{t, \tau, A} = e^{\top} \left( \int^t_\tau \log (\lambda^*_s/\lambda_s) dN_v - \int^t_\tau (\lambda^*_s - \lambda_s) ds \right). 
\]  
where \(./\) denotes elementwise division. We find a change-point whenever \(\sup_{t, \tau, A} \ell_{t, \tau, A} > x\).

### 4.2 Changes from Hawkes Process to Hawkes Process with different parameters

Next, we construct the hypothesis test to detect the intensity changes where the reference data follow Hawkes processes. The changes in intensity would be reflected in the changes in influence parameter \(\alpha\) or \(A\). Assume that the Hawkes processes are stationary. We check the data stored in time interval \((\tau, t)\), and let the time window \(t - \tau\) is relatively large. Based on the above assumptions, suppose \(\tau\) is a change-point, we construct the hypothesis test as
\[
H_0: \quad \lambda_s = \mu + \alpha \int^s_\tau \varphi(s-v) dN_v, \quad 0 < s < t \\
H_1: \quad \lambda^*_s = \mu + \alpha^* \int^s_\tau \varphi(s-v) dN_v, \quad 0 < s < \tau \\
\lambda^*_s = \mu + \alpha^* \int^s_\tau \varphi(s-v) dN_v, \quad \tau \leq s < t. 
\]  

We can correspondingly write out the log-likelihood ratio as (5). Plug in the expressions for \( \lambda_s \) and \( \lambda^*_s \), adopt exponential kernel \( \varphi(t) = \beta e^{-\beta t} \), and define \( L = t - \tau \), we finally obtain

\[
\ell_{t,\tau,\alpha^*} = \sum_{i=1}^{n} \log \left[ \frac{\mu + \alpha^* \sum_{j=1}^{i-1} \beta e^{-\beta(t_i-t_j)}}{\mu + \alpha \sum_{j=1}^{i-1} \beta e^{-\beta(t_i-t_j)}} \right] \\
- (\alpha^* - \alpha) \sum_{i=1}^{n} \left[ 1 - e^{-\beta(t-t_i)} \right] .
\]

(11)

Still, \( \alpha^* \) can be estimated from data between time \((\tau, t)\). And we detect a change-point whenever \( \sup_{t,\tau,\alpha^*} \ell_{t,\tau,\alpha^*} > x \).

For the high dimensional case, we need to express the hypothesis test in matrix form

\[
H_0 : \lambda_s = \mu + A \int_{\tau}^{s} \varphi(s-v) dN_v, \ \tau < s < t; \\
H_1 : \lambda^*_s = \mu + A^* \int_{\tau}^{s} \varphi(s-v) dN_v, \ \tau < s < t.
\]

(12)

Suppose the total number of events occurred in between \((\tau, t)\) is \( n \), and we record all the events as \( \{(t_1, u_1), (t_2, u_2), \ldots, (t_N, u_N)\} \), where \( u_i \in \{1, 2, \ldots, d\} \) indicates the location of the event. The specific log-likelihood ratio for this setting is:

\[
\ell_{t,\tau,A^*} = \sum_{i=1}^{n} \log \left[ \frac{\mu_{u_i} + \sum_{j \leq t_i} \alpha^*_{u_i,u_j} \beta e^{-\beta(t_i-t_j)}}{\mu_{u_i} + \sum_{j \leq t_i} \alpha_{u_i,u_j} \beta e^{-\beta(t_i-t_j)}} \right] \\
- \sum_{j=1}^{d} \sum_{i=1}^{n} (\alpha^*_{j,u_i} - \alpha_{j,u_i}) \left( 1 - e^{-\beta(t-t_i)} \right) .
\]

(13)

Remark. The topology of the network has been embedded in the sparsity pattern of the influence matrix \( A \), which are given a priori. The values in \( A \) are estimated online. The dependency between different nodes in the network and the temporal dependence over events can be fully characterized in the process of updating (or tracking) the influence matrix \( A \) as events stream in. This can be easily achieved as an online convex optimization problem with warm start (i.e., we always initialize the parameters using the optimal solutions of the last step).

4.3 EM-like algorithm for online estimation

Even though there does not exist a closed-form estimator for influence parameter \( \alpha \) or influence matrix \( A \), we can easily update these parameters online by solving a convex optimization problem via EM-like algorithm. Suppose \( L = t - \tau \) is a large number. We can estimate \( \alpha \) based on the events within time interval \((\tau, t)\). For simplicity of exposition, we will focus on one dimensional case. We want to find an estimator \( \hat{\alpha} \) which maximize the likelihood
Figure 2: Illustration of online algorithm.

function:

\[ L_{t, \tau, \alpha} = \sum_{i=1}^{n} \log \left( \mu + \alpha \sum_{\tau < t_j < t_i} \beta e^{-\beta(t_i - t_j)} \right) - \mu L \]

\[ - \alpha \sum_{i=1}^{n} \left[ 1 - e^{-\beta(t_i - t_j)} \right], \]

which is a concave function w.r.t. \( \alpha \). One can use gradient descent to optimize this objective.

However, we want to obtain an algorithm free of tuning parameters. We will first use Jensen’s inequality to obtain a concave lower bound by introducing auxiliary variables \( p_{ij} \) for all pair of events

\[ \sum_{j=1}^{i} p_{ij} = 1; \quad p_{ij} \geq 0; \quad 1 \leq j \leq i \leq n \]

such that

\[ L_{t, \tau, \alpha} \geq \sum_{i=1}^{n} \left( p_{i} \log \mu + \sum_{j=1}^{i-1} p_{ij} \log \left[ \alpha \sum_{\tau < t_j < t_i} \beta e^{-\beta(t_i - t_j)} \right] \right) \]

\[ - \sum_{j=1}^{i} p_{ij} \log p_{ij} \right) - \mu L - \alpha \sum_{i=1}^{n} \left[ 1 - e^{-\beta(t_i - t_j)} \right], \]

where \( \{p_{ij}\} \) can be interpreted as the probability that event \( j \) triggers event \( i \). Note that the lower-bound is valid for every choice of \( \{p_{ij}\} \) which satisfies (15). To make the lower bound tight and ensure improvement in each iteration, we will maximize it with respect to \( p_{ij} \) and obtain (assuming we have \( \alpha^{k} \) from previous iteration or initialization). Once we have the tight lower bound, we will take gradient of this lower-bound with respect to \( \alpha \). We iterate these two steps:

E-step: \[ p_{ij}^{k} = \frac{\alpha^{k} \beta e^{-\beta(t_j - t_i)}}{\mu + \sum_{i=1}^{j-1} \alpha^{k} \beta e^{-\beta(t_j - t_i)}}, 1 \leq j < i \leq n. \] (16)

M-step: \[ \alpha^{k+1} = \frac{\sum_{i<j} p_{ij}^{k}}{n - \sum_{j=1}^{n} e^{-\beta(t_j - t_j)}}, \] (17)

until the algorithm converges and obtain the estimated \( \alpha \). In practice, we find that we only need 3 or 4 iterations to obtain good estimation if using warm start, which we will use in our experiments.
Similarly, for high-dimensional case, $A$ is estimated via:

**E-step:**  
$$p^{k}_{ij} = \frac{\alpha_{u_i,u_j}^k \beta e^{-\beta(t_i-t_j)}}{\mu_{u_i} + \sum_{j=1}^{i-1} \alpha_{u_i,u_j}^k \beta e^{-\beta(t_i-t_j)}}, 1 \leq j < i \leq n; \quad (18)$$

**M-step:**  
$$\alpha_{u,v}^{k+1} = \frac{\sum_{i: u_i = u} \sum_{j < i: u_j = v} p^{k}_{ij}}{\sum_{j: u_j = v} (1 - e^{-\beta(t-t_j)})}. \quad (19)$$

Overall, our algorithm can be summarized as Fig. [2]. We fix the length of the time window to be $L = t - \tau$. The user should set the window length $L$ beforehand when implementing the algorithm. A proper length of window relies on the nature of the data. If the data are noisy, usually a longer time window is needed to have a better estimation of the parameter and reduce the false alarm. We move the window forward every new event or every $\gamma$ events. Every time we move the sliding window we will estimate the influence parameters $\hat{\alpha}$ based on the events within the sliding window and construct the log-likelihood ratio as discussed beforehand. In addition, we can leverage the warm start trick when we perform the online estimation. For two adjacent sliding window, the optimal solution ($\hat{\alpha}$) of a later window should be very close to the previous one. We therefore use the optimal solution from the previous time window as the initialized parameter for the next time window, which greatly increase the efficiency of our algorithm. We claim a change-point with location $\tau$ and stop the algorithm whenever the log-likelihood ratio exceeds a predefined threshold.

## 5 Theoretical threshold

A key step in implementing the online detection algorithm is to set the threshold. The choice of threshold is a trade-off between two standard performance metrics: (i) the expected stopping time when there is no change-points, or named average run length (ARL); and (ii) the expected detection delay when there exists a change-point. Usually, the threshold is estimated via direct Monte Carlo by relating threshold to ARL assuming the data follow the null distribution. However, Monte Carlo is not only computationally expensive, in some practical problems, repeated experiments would be prohibitive. Therefore it is essential to find a cheaper way to accurately estimate the threshold.

We develop an analytical function which relates the threshold to ARL. That is, given a prescribed ARL, we can solve for the corresponding threshold $x$ analytically. We put the key theorem below. The main techniques we are using are change-of-measure, mean field approximation for point processes, and Delta method.

In our algorithm, we fix the time window $L = t - \tau$, and stop whenever the log-likelihood ratio $\ell_{t,\tau,\alpha}$ exceeds the pre-defined threshold $x > 0$ for the first time. The stopping time is defined as

$$T = \inf \left\{ t : \sup_{\alpha} \ell_{t,\tau,\alpha} > x \right\}. \quad (20)$$

Therefore the ARL can be written as $\mathbb{E}^\infty[T]$, where $\mathbb{E}^\infty$ denotes the expectation under the null distribution.

**Theorem 1** (ARL under the null distribution). When $x \to \infty$ and $x/\sqrt{L} \to c'$ for some constant $c'$, the average run length (ARL) of the stopping time $T$ defined in (20), for one-
dimensional case,
\[
\mathbb{E}^\infty[T] = e^x \left[ \int_{\alpha \in \Theta} \nu \left( \frac{2\xi}{\eta^2} \right) \phi \left( \frac{\ell t,\tau,\alpha}{\sqrt{L\sigma^2}} \right) d\alpha \right]^{-1} \cdot [1 + o(1)];
\]
for high-dimensional case, the same expression holds except that \( \int_{\alpha} \) is replaced by \( \int_{\mathcal{A}} \), which means taking integral with respect to all the nonzero entries of the matrix \( \int_{\mathcal{A}} = \int \cdots \int_{(\alpha_{ij}, \alpha_{ij} \neq 0)} \).

Above,
\[
\nu(\mu) = \frac{(2/\mu) (\Phi(\mu/2) - 0.5)}{(\mu/2) \Phi(\mu/2) + \phi(\mu/2)},
\]
\( I = \mathbb{E}_{t,\tau,\alpha}[\ell t,\tau,\alpha]/L, \quad \sigma^2 = \text{Var}_{t,\tau,\alpha}[\ell t,\tau,\alpha]/L, \)
\( \xi = -(\mathbb{E}[\ell t,\tau,\alpha] - \mathbb{E}_{t,\tau,\alpha}[\ell t,\tau,\alpha])/L, \)
\( \eta^2 = (\text{Var}[\ell t,\tau,\alpha] + \text{Var}_{t,\tau,\alpha}[\ell t,\tau,\alpha])/L. \)

Here \( \Phi(x) \) is the cumulative distribution function of the standard normal; \( \phi(x) \) is the probability density function of the standard normal; \( \mathbb{E}[\ell t,\tau,\alpha] \) and \( \text{Var}[\ell t,\tau,\alpha] \) denote the expectation and the variance of the log-likelihood ratio under the null distribution (i.e., suppose the data follow \( H_0 \)) and \( \mathbb{E}_{t,\tau,\alpha}[\ell t,\tau,\alpha] \) and \( \text{Var}_{t,\tau,\alpha}[\ell t,\tau,\alpha] \) denote the expectation and variance under the alternative distribution (i.e., suppose the data follow \( H_1 \)). If we further define \( I_0 = \mathbb{E}[\ell t,\tau,\alpha]/L \) and \( \sigma_0^2 = \text{Var}[\ell t,\tau,\alpha]/L \), all the calculations will be finally boiled down to the evaluations of \( I, I_0, \sigma^2, \) and \( \sigma_0^2 \). Since we construct the log-likelihood ratio based on two scenarios: data change from Poisson process to Hawkes process; data change from Hawkes process to another Hawkes process with different influence parameters. We summarize the computing results for \( I, I_0, \sigma^2, \) and \( \sigma_0^2 \) under all these settings in Table 1. For some notations undefined in the table, we add them here:
\[
\hat{\lambda}^* = (I - A^*)^{-1} \mu, \quad \bar{\lambda} = (I - A)^{-1} \mu,
\]
\[
H = [\log ((I - A)^{-1} \mu) - \log (\mu)] [\log ((I - A)^{-1} \mu) - \log (\mu)]^T,
\]
\[
C = (I - A)^{-1} A (2I + (I - A)^{-1} A) \text{ diag } ((I - A)^{-1} \mu)
+ \text{ diag } ((I - A)^{-1} \mu),
\]
\[
G_{ij} = \log \left( \frac{\lambda_i}{\lambda_j} \right) \log \left( \frac{\lambda_i}{\lambda_j} \right), \quad F_{ij} = \left( 1 - \frac{\lambda_i}{\lambda_j} \right) \left( 1 - \frac{\lambda_i}{\lambda_j} \right),
\]
\[
R_{ij} = \left( \frac{\lambda_i}{\lambda_i} - 1 \right) \left( \frac{\lambda_j}{\lambda_j} - 1 \right), \quad 1 \leq i \leq j \leq d.
\]
\[
C^* = (I - A^*)^{-1} A^* (2I + (I - A^*)^{-1} A^*) \text{ diag } ((I - A^*)^{-1} \mu)
+ \text{ diag } ((I - A^*)^{-1} \mu),
\]
We will evaluate the accuracy of Theorem 1 in section 6.3.
Table 1: Computations of $I$, $I_0$, $\sigma^2$ and $\sigma_0^2$ under different settings.

| Setting                        | $I$                                    | $I_0$                              | $\sigma^2$                      | $\sigma_0^2$                      |
|-------------------------------|----------------------------------------|------------------------------------|----------------------------------|-----------------------------------|
| Poi. $\rightarrow$ Haw. (one dim.) as shown in [2] and [5] | $\frac{\mu}{1-\alpha} \log \left( \frac{1}{1-\alpha} \right) - \frac{\alpha}{1-\alpha} \mu$ | $\mu \log \left( \frac{1}{1-\alpha} \right) - \frac{\alpha}{1-\alpha} \mu$ | $\left[ \log \left( \frac{1-\alpha}{1-\alpha} \right) \right]^2$ | $\mu \left[ \log \left( \frac{1-\alpha}{1-\alpha} \right) \right]^2$ |
| Poi. $\rightarrow$ Haw. (high dim.) as shown in [8] and [9] | $\lambda^\top \left( \log(\lambda^*) - \log(\mu) \right)$ | $-e^\top (\lambda^* - \mu)$ | $\left[ \log \left( \frac{1-\alpha}{1-\alpha} \right) \right]^2 \cdot$ | $1 - \frac{1-\alpha}{1-\alpha} \mu$ |
| Haw. $\rightarrow$ Haw. (one dim.) as shown in [10] and [11] | $\frac{\mu}{1-\alpha} \log \left( \frac{1}{1-\alpha} \right) - \frac{\alpha}{1-\alpha} \mu$ | $-e^\top (\lambda^* - \mu)$ | $\mu \left[ \log(\lambda^*) - \log(\mu) \right]$ | $\mu^\top (\log(\lambda^*) - \log(\mu))$ |
| Haw. $\rightarrow$ Haw. (high dim.) as shown in [12] and [13] | $\lambda^\top \left[ \log(\lambda^*) - \log(\lambda) \right]$ | $-e^\top (\lambda^* - \lambda)$ | $\left[ \log \left( \frac{1-\alpha}{1-\alpha} \right) \right]^2 \cdot$ | $\mu \left[ \log(\lambda^*) - \log(\mu) \right]$ | $\mu^\top (\log(\lambda^*) - \log(\mu))$ |

6 Experiments

In this section, we conducted experiments using both synthetic and real-world dataset. We will focus on our experiments on the efficiency of the statistic, comparison with baseline, the accuracy of the analytic threshold and illustration with real-world data.

6.1 Detection efficiency of statistic

We consider various scenarios and demonstrate the detection statistic for each case. The results are shown in Fig. 3. We set the ARL to be 100000 and obtain the corresponding thresholds under different cases via direct Monte Carlo. In all the example, thresholds are plotted as red lines and the dashed green lines indicate the true change-point location.

In Fig. 3(a), we consider a situation when the events first follow a one-dimensional Poisson process with intensity $\mu = 1$ then shift to a Hawkes process with influence parameter $\alpha = 0.3$ and $\beta = 1$. When applying our algorithm, we choose the window length $L = 100$. The true change-point occurs at time 1000. In this example, we see the statistic computed by our algorithm behaves as a flat line (the actual value ranges from $10^{-11}$ to $10^{-6}$) when there is no change-point. This indicates our method is very robust when there is no change-point, which would lead to a lower false alarm. On the other hand, when the change-point appears, our statistic responds quickly, shows a sharp jump and touches the threshold. Some reasonable time delay is unavoidable since we introduce sliding window and more post-data are needed to be involved in the sliding window to have a better estimation of the influence parameter.

In Fig. 3(b), we consider a hard case when the data transit from one Hawkes process with parameter values $\mu = 1$, $\alpha = 0.3$ and $\beta = 10$ to another Hawkes process with a larger influence parameter $\alpha = 0.5$. The true change-point is around time 1000 and the window length we use is $L = 100$. Still, our statistic is extremely powerful and detect such a change as soon as it emerges.

In Fig. 3(c), we will consider high-dimensional cases. The pre-change data are two-dimensional i.i.d. Poisson processes each with intensity 0.5. Then from time 1000, the self- and mutual-excitation appear and the processes change to two-dimensional Hawkes processes.

11
with influence matrix $A = [0.1, 0.3; 0.2, 0.1]$ and $\beta = 1$. We set $L = 400$.

In Fig. 3(d), the pre-change data are two-dimensional Hawkes processes, with base intensity $\mu = [0.3, 0.3]^{\top}$, $A = [0.1, 0.3; 0.2, 0.1]$ and $\beta = 1$. After the change-point, the influence matrix becomes $A = [0.5, 0.4; 0.4, 0.3]$. We choose $L = 400$. We want to emphasize that for the high-dimensional case especially for the high-frequency data, there is no need to update the influence parameter given a new event. We can choose a resolution according to the specific cases. For example, in both the third and the fourth example, we only update the parameters given 10 new events.

In Fig. 3(e), we consider a network with one parent and nine children (star shape), which is commonly used to model how the information broadcasting over the network. Before the change-point, each note has a base intensity 0.1 and the influence from the parent to each child is 0.3. After the change-point, this influence increases to 0.5. Since the base intensity is very low and such a change for each note should be very weak and hard to detect. But our statistic can automatically aggregate such a change based on topology.

In Fig. 3(f), we consider a network with a chain of 10 nodes, which is commonly used to model how the information propagation over the network. The parameter settings are the same with the star example. For these two example, we select the window length $L = 400$ and let $\beta = 1$. In the last example, we assume the network topology is generic.

Actually, our proposed algorithm can work for network with any topology. In Fig. 3(f), we generate a Erdős-Rényi random network consisting of 10 nodes. The parameter setting is the same with the star and chain case. Obviously, our algorithm also work for general network structure.
6.2 Comparison with baseline

We compare our proposed statistic to the traditional log-likelihood ratio based on Poisson and \textit{i.i.d.} assumptions in a discrete time setting. For this baseline approach, we need to discretize the time into bins and count the number of events within each bin.

Suppose \(n_1, n_2, \ldots, n_c\) are the sequence of counting numbers following Poisson distribution with intensity \(\lambda_i\), \(i = 1, 2, \ldots, c\) is the index of the discrete time step. The baseline method constructs the hypothesis test as:

\[
\begin{align*}
H_0: & \quad \lambda_i = \mu, \quad i = 1, 2, \ldots, c; \\
H_1: & \quad \lambda_i = \mu, \quad i = 1, 2, \ldots, k; \\
\lambda_i^* = \mu^*, & \quad i = k + 1, \ldots, c,
\end{align*}
\]

where \(k\) is a potential change-point. We can then write out the likelihood function under the null hypothesis as (use one-dimensional case as an example)

\[
\prod_{i=1}^c e^{-\mu n_i} \frac{\mu^{n_i}}{n_i!}.
\]

The log-likelihood function \(L_{H_0}\) is given by

\[-c\mu + \sum_{i=1}^c n_i \log \mu - \sum_{i=1}^c \log n_i!,\]

and the likelihood function under the alternative function is given by

\[
\prod_{i=1}^k e^{-\mu n_i} \frac{\mu^{n_i}}{n_i!} \prod_{i=k+1}^c e^{-\mu^* n_i} \frac{\mu^*^{n_i}}{n_i!}.
\]

Hence, the log-likelihood function \(L_{H_1}\) is given by

\[-k \mu - (c-k) \mu^* + \sum_{i=1}^c n_i \log \mu + \sum_{i=k+1}^c n_i \log \mu^* - \sum_{i=1}^c \log n_i!.
\]

Finally, we obtain the log-likelihood ratio statistic as

\[
\ell_{k,c,\mu^*} = -(c-k)(\mu^* - \mu) + \sum_{i=k+1}^c n_i \log \frac{\mu^*}{\mu}.
\]

We can also estimate \(\mu\) from reference data and replace \(\mu^*\) by its maximum likelihood estimator \(\hat{\mu}^*\), which is given by \(\hat{\mu}^* = \sum_{i=k+1}^c n_i / (c-k)\). We claim a change-point with location \(k\) whenever \(\sup_{k,c,\mu^*} \ell_{k,c,\mu^*} > x\), where \(x\) is a pre-determined threshold. For the baseline approach, we assume every dimension follows an independent Poisson process. Therefore, the log-likelihood ratio for the high-dimensional case is just a summation of the log-likelihood ratio for each dimension. Suppose the total dimension is \(d\), we have

\[
\ell_{k,c,\mu^*} = \sum_{j=1}^d \left[-(c-k)(\mu_j^* - \mu_j) + \sum_{i=k+1}^c n_i^j \log \frac{\mu_j^*}{\mu_j}\right].
\]

We compare our approach to the above-mentioned baseline on various synthetic datasets. For each dataset, we created 500 event sequences with half of them containing one change-point and half of them containing no change-points. We then plot the AUC curves (true positive rate v.s. false positive rate under various threshold) for the two approaches on each dataset, as shown in Fig. 4.

Here are some brief descriptions of the synthetic datasets. In Fig. 4 A.1–A.4, the data
before change-points follow one-dimensional Poisson process and the post-change data follow Hawkes processes with different influence parameters. In Fig. 4 B.1-B.3, the pre- and post-change data both follow Hawkes processes with the same base intensity but with different influence parameters. In Fig. 4 C.1 to C.2, the data before change-points are two dimensional i.i.d. Poisson processes and then the self- and mutual-excitation appears. In Fig. 4 D.1-D.3, the pre- and post-change data are intertwined Hawkes processes with star topology. Then the levels of the mutual-excitation shifts. For both the two approaches, we use the same length of sliding window for a fair comparison for every dataset. For our approach, we simply use all the events within the time window to compute the statistic as discussed beforehand. For the baseline, we discretize the time window into bins and count the number of events within each bin and compute the statistic as (22).

The detailed parameter settings for all the examples are listed below Fig. 4. We only put some key findings here. Our approach has a distinctively better discriminative performance compared to baseline on all datasets. Our method has a higher detection rate for the true change-points and a lower false alarm. Especially for weak signals, the baseline approach is just slightly better than the random guess (would correspond to the diagonal line), whereas our approach consistently works well. In addition, our approach has a pronounceable superior performance especially under large $\beta$. One possible explanation to this phenomenon is that the EM-like algorithm tends to have a more accurate, less variance estimation of the influence parameter $\alpha$ when $\beta$ is large. And it also converges faster. This is discovered by [14]. Since our detection algorithm relies on the estimation of the influence parameter $\alpha$, a more accurate estimation would definitely enhance the discriminative power.

### 6.3 Accuracy of theoretical threshold

We evaluate the accuracy of the approximated theoretical threshold discussed in Theorem 1 by comparing with the true threshold obtained by direct Monte Carlo. We consider various scenarios and tune the parameter settings. We demonstrate the results in Fig. 5 and list the parameter settings here.

For Fig. 5(a)(b)(c), the reference data is one-dimensional Poisson process with intensity $\mu = 1$. We choose $\beta = 1$ as a priori when online computing the statistic. The only difference is the length of the sliding time window. We set $L = 10$, $L = 50$ and $L = 100$ respectively. For Fig. 5(d), we select the sliding time window to be $L = 50$ but let $\beta = 10$. By comparing these four examples, we claim our approximated threshold is very accurate regardless of how long the time window we choose and what the prior parameter $\beta$ we set.

For Fig. 5(e)(f), the reference data is one-dimensional Hawkes process with base intensity $\mu = 1$ and influence parameter $\alpha = 0.3$, $\beta = 10$. We tune the sliding time window to be $T = 100$ and $T = 150$ respectively. We can obtain the same conclusions here. For Fig. 5(g)(h), we consider a high-dimensional case. The reference data follow two dimensional i.i.d. Poisson processes with base intensity $\mu = [0.5, 0.5]^\top$. We set $\beta = 1$ and tune the sliding window to be $L = 300$ and 400 respectively. The results demonstrate that our analytical threshold is also sharply accurate in the high-dimensional situation.
Figure 4: AUC curves: comparison with baseline. For A.1–A.3, pre-change data follow one dimensional Poisson process with $\mu = 1$; post-change data follow one dimensional Hawkes process with $\alpha = 0.2$, $\beta = 1,10,100$ respectively; for A.4, $\alpha = 0.3$ and $\beta = 10$. For B.1–B.3, pre-change data follow one dimensional Hawkes process with $\alpha = 0.3$; post-data with $\alpha = 0.5$; $\beta = 1,10,100$ respectively. For C.1–C.2, pre-data follow two dimensional i.i.d. Poisson processes with $\mu = [0.2,0.2]^T$; post-data follow two dimensional Hawkes processes with extra influence parameter $A = [0.1,0.1;0.1,0.1]; \beta = 1,10$ respectively. For D.1, before change-point, for each node $\mu = 0.1$, $\beta = 1$, the influence from parent to each child is $\alpha = 0.3$; after change-point $\alpha = 0.4$; for D.2–D.3, $\alpha$ changes to 0.5, $\beta = 1,10$ respectively.

7 Real-data

We evaluated our online detection algorithm on real Twitter and news websites data. The twittering behaviors over the network, for example, can be modeled as multi-dimensional point processes. By evaluating our proposed log-likelihood ratio just based on the real twittering events, we found that the statistics would rise up when there is an explanatory major event in actual scenario. By comparing the detected change points to the true major event time, we verify the accuracy and effectiveness of our proposed algorithm. In all our real experiments, we set the sliding window size to be $L = 500$ minutes, we compute our statistic per 10 events and set the kernel bandwidth $\beta$ to be 1. The number of events tested ranges from 3000 to 15000 for every dataset.
Figure 5: Comparison of theoretical threshold with simulated threshold.

Figure 6: Exploratory results on Twitter for the detected change-points: (a) Mr Robot wins the Golden Globe; (b) First Lady’s dress getting attention; (c) Suresh Raina makes his team won; (d) Court hearing on Martin Shkreli; (e) Rihanna listens to ANTI; (f) Daughter releases his new album.

Figure 7: Exploratory results on Memetracker for the detected change-points: (a) Obama wins the presidential election; (b) Israel announces ceasefire; (c) Beijing Olympics starts.
Twitter dataset. For Twitter dataset we focus on the star network topology. We created a dataset for 6 famous people users and randomly select about 30 of their follower among the tens of thousands followers. We assume there is a star-shaped influence network from the celebrity to the followers. We collect all their re/tweets in late January and early February 2016.

Figure 6-a demonstrates the statistic computed for the account associated to a TV series named Mr. Robot. We identified that the statistic increases around late January 10th and early 11th. This, surprisingly corresponds to the winning of the 2016 Golden Glob Award.

Figure 6-b shows the statistic computed based on the events of the First lady of the USA and 30 of her randomly selected followers. This statistic revealed a sudden increase in 13th of January. We found a related event - Michelle Obama stole the show during the president’s final State of the Union address by wearing a marigold dress which sold out even before the president finished the speech. This event obviously changed the twittering behaviors of the community consisting of the first lady and a couple her followers.

Figure 6-c is related to Suresh Raina, an Indian professional cricketer. We selected a small social circle around him as the center of a star-shaped influence network. We noticed that he led his team to win an important game on Jan. 20, which corresponds to a sharp increase of the computed statistic.

The scenario for Figure 6-d is also interesting as it reflects the activity on the network surrounding Mr. Shkreli, the former chief executive of Turing Pharmaceuticals, who is facing federal securities fraud charges. At Feb. 4th he was invited to congress for a hearing to be questioned about drug price hikes. The 5th example is about Rihanna who announced the release of her new album in a tweet on Jan. 25th. That post was retweeted 170K times and received 280K likes and creates a sudden change in network of her followers.

2 http://www.tvguide.com/news/golden-globe-awards-winners-2016/
3 http://www.cnn.com/2016/01/13/living/michelle-obama-dress-marigold-narciso-rodriguez-feat/
4 http://www.espncricinfo.com/syed-mushtaq-ali-trophy-2015-16/content/story/963891.html
5 http://www.nytimes.com/2016/02/05/business/drug-prices-valeant-martin-shkreli-congress.html
6 http://jawbreaker.nyc/2016/01/is-rihannas-anti-album-finally-done/
The last example, in Figure 6-f, demonstrates an increase in the statistic related to the network of Daughter around 25th of January who is attributed to releasing his new album at Jan. 25th.

**Memetracker dataset.** We also analyzed the well-known dataset Memetracker\(^8\) for a specific question.

First we selected the news websites which mentioned the word “Obama” in their memes during 2nd to 5th of November 2008. We extracted top 70 web domains based on the frequency and construct an underlying influence network based on the reported linking information available within the dataset. This subset always includes the credible news agencies such as BBC, CNN, Huffingtonpost, Guardian, etc. By tracking the event data we successfully pinpoint a change right at the time that Barak Obama was elected as the 44th president of the USA.\(^9\) This is illustrated in Figure 7-a. We also plotted the topology of the network that we used, as shown in Figure 8.

In another example we found a sharp change in the statistic computed based on the data of a group which contains 70 websites talking about Israel in their memes from 13th to 18th of January. As demonstrated in Figure 7-b, the statistic increases right at the time the Israel announces a unilateral ceasefire in the Gaza War back in 2009.\(^10\)

Figure 7-c is associated to 2008 Summer Olympics in Beijing, China. The change-point detected is exactly the starting day, August 8th.\(^11\)

**References**

[1] E. Bacry and J.-F. Muzy. Second order statistics characterization of hawkes processes and non-parametric estimation. arXiv preprint arXiv:1401.0903, 2014.

[2] N. A. Christakis and J. H. Fowler. Social network sensors for early detection of contagious outbreaks. PloS one, 5(9):e12948, 2010.

[3] D. J. Daley and D. Vere-Jones. Scoring probability forecasts for point processes: the entropy score and information gain. Journal of Applied Probability, pages 297–312, 2004.

[4] D. J. Daley and D. Vere-Jones. An introduction to the theory of point processes: volume II: general theory and structure. Springer Science & Business Media, 2007.

[5] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In Advances in Neural Information Processing Systems (NIPS), 2014.

[6] M. Farajtabar, M. Gomez-Rodriguez, N. Du, M. Zamani, H. Zha, and L. Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In

---

\(^7\)http://www.nme.com/news/daughter/79540

\(^8\)http://www.memetracker.org/

\(^9\)https://en.wikipedia.org/wiki/United_States_presidential_election,_2008

\(^10\)http://news.bbc.co.uk/2/hi/middle_east/7835794.stm

\(^11\)https://en.wikipedia.org/wiki/2008_Summer_Olympics
[7] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *NIPS ’15: Advances in Neural Information Processing Systems*, 2015.

[8] E. W. Fox, M. B. Short, F. P. Schoenberg, K. D. Coronges, and A. L. Bertozzi. Modeling e-mail networks and inferring leadership using self-exciting point processes. *Submitted to Journal of the American Statistical Association*, 2014.

[9] E. C. Hall and R. M. Willett. Tracking dynamic point processes on networks. *arXiv preprint arXiv:1409.0031*, 2014.

[10] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[11] T. Herberts and U. Jensen. Optimal detection of a change point in a poisson process for different observation schemes. *Scandinavian Journal of Statistics*, 31(3):347–366, 2004.

[12] N. Laptev, S. Amizadeh, and I. Flint. Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1939–1947. ACM, 2015.

[13] C. Lévy-Leduc and F. Roueff. Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, pages 637–662, 2009.

[14] E. Lewis and G. Mohler. A nonparametric em algorithm for multiscale hawkes processes. *preprint*, pages 1–16, 2011.

[15] S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*, 2014.

[16] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai. Distinguishing infections on different graph topologies. *Information Theory, IEEE Transactions on*, 61(6):3100–3120, 2015.

[17] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web*, pages 913–924. ACM, 2014.

[18] J. Neyman and E. S. Pearson. *On the problem of the most efficient tests of statistical hypotheses*. Springer, 1992.

[19] J. C. L. Pinto, T. Chahed, and E. Altman. Trend detection in social networks using hawkes processes. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1441–1448. ACM, 2015.
[20] S. Rajaram, T. Graepel, and R. Herbrich. Poisson-networks: A model for structured point processes. In Proceedings of the 10th international workshop on artificial intelligence and statistics, pages 277–284. Citeseer, 2005.

[21] R. Ratnam, J. Goense, and M. E. Nelson. Change-point detection in neuronal spike train activity. Neurocomputing, 2003.

[22] M. G. Rodríguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. arXiv preprint arXiv:1105.0697, 2011.

[23] Y. Saatçi, R. D. Turner, and C. E. Rasmussen. Gaussian process change point models. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 927–934, 2010.

[24] J. J. Shen, N. R. Zhang, et al. Change-point model on nonhomogeneous poisson processes with application in copy number profiling by next-generation dna sequencing. The Annals of Applied Statistics, 6(2):476–496, 2012.

[25] V. Solo and A. Pasha. A test for independence between a point process and an analogue signal. Journal of Time Series Analysis, 33(5):824–840, 2012.

[26] D. Vere-Jones. Probabilities and information gain for earthquake forecasting. Selected Papers From Volume 30 of Vychislitel’naya Seysmologiya, pages 104–114, 1998.

[27] B. Yakir. Extremes in random fields: a theory and its applications. John Wiley & Sons, 2013.

[28] S.-H. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 1–9, 2013.

[29] N. R. Zhang, B. Yakir, C. L. Xia, and D. Siegmund. Scanning a poisson random field for local signals. arXiv preprint arXiv:1406.3258, 2014.

[30] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 1301–1309, 2013.

[31] J. R. Zipkin, F. P. Schoenberg, K. Coronges, and A. L. Bertozzi. Point-process models of social network interactions: parameter estimation and missing data recovery. European Journal of Applied Mathematics, pages 1–28, 2015.
A Proofs of Theorem 1

A.1 Overall framework in deriving Theorem 1

We use the one dimensional case as an example. Let $t$ be the latest time, fix time window $L = t - \tau$, we want to evaluate the extreme value of the probability $\mathbb{P}\{\sup_{t, \alpha} \ell_{t, \tau, \alpha} > x\}$ under the null distribution within the following parameter regions:

\begin{align}
    t &< m \quad (23) \\
    \tau &= t - L \quad (24) \\
    \alpha &\in \Theta, \quad \Theta = (0, 1) \quad (25)
\end{align}

Note that all the parameter spaces are smooth. In the following we will use the “likelihood ratio identity” trick, which computes a probability of an event formulated in some distribution by reformulating it as an expectation in the context of an alternative distribution. We use the notation $\mathbb{P}_{t, \tau, \alpha}$ to refer to the probability measure under the alternative distribution and use the notation $\mathbb{E}_{t, \tau, \alpha}$ to indicate the expectation in the context of the alternative distribution. Then we will have:

\begin{align}
    \mathbb{P}\left[ \sup_{t, \alpha} \ell_{t, \tau, \alpha} > x \right] &= \mathbb{E}\left[ \left\{ \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right\} \right] = \mathbb{E}\left[ 1; \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right]. \quad (26)
\end{align}

We are using the notation $\mathbb{E}[U; A] = \mathbb{E}[U \mathbb{1}\{A\}]$ in the last equality. Then, we have:

\begin{align}
    \mathbb{P}\left[ \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] &= \mathbb{E}\left[ \frac{\int_t \int_{\alpha \in \Theta} e^{\ell_{t, \tau, \alpha} dt} d\alpha}{\int_t' \int_{\alpha' \in \Theta} e^{\ell_{t', \tau', \alpha'} dt'} d\alpha'}; \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] \quad (27) \\
    &= \int_t \int_{\alpha \in \Theta} \mathbb{E}\left[ \frac{e^{\ell_{t, \tau, \alpha}}}{\int_t' \int_{\alpha' \in \Theta} e^{\ell_{t', \tau', \alpha'} dt'} d\alpha'}; \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] dt d\alpha \quad (28) \\
    &= \int_t \int_{\alpha \in \Theta} \mathbb{E}_{t, \tau, \alpha}\left[ \frac{1}{\int_t' \int_{\alpha' \in \Theta} e^{\ell_{t', \tau', \alpha'} dt'} d\alpha'}; \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] dt d\alpha \quad (29)
\end{align}

where the second equality follows from changing the order of summation and the expectation, and for the last equality, we are using the change-of-measure trick

\begin{align}
    d\mathbb{P} = e^{-\ell_{t, \tau, \alpha}} d\mathbb{P}_{t, \tau, \alpha}. \quad (30)
\end{align}

Next, we have:

\begin{align}
    \mathbb{P}\left[ \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] &= e^{-x} \int_t \int_{\alpha \in \Theta} \mathbb{E}_{t, \tau, \alpha}\left[ \frac{1}{\int_t' \int_{\alpha' \in \Theta} e^{\ell_{t', \tau', \alpha'} dt'} d\alpha'}; \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] dt d\alpha. \quad (31)
\end{align}

We rearrange each term and add notations:

\begin{align}
    \mathbb{P}\left[ \sup_{t, \alpha \in \Theta} \ell_{t, \tau, \alpha} > x \right] &= e^{-x} \int_t \int_{\alpha \in \Theta} \mathbb{E}_{t, \tau, \alpha}\left[ \frac{\mathcal{M}_{t, \tau, \alpha}}{S_{t, \tau, \alpha}} e^{-[\ell_{t, \tau, \alpha} + m_{t, \tau, \alpha}]; \ell_{t, \tau, \alpha} + m_{t, \tau, \alpha} > 0} \right] dt d\alpha. \quad (32)
\end{align}
where

\[ M_{t,\tau,\alpha} = \sup_{t'} e^{t',\alpha'} - \ell_{t,\tau,\alpha} \]

\[ S_{t,\tau,\alpha} = \int_{t'} e^{t',\alpha'} - \ell_{t,\tau,\alpha} \, dt' \]

\[ \bar{\ell}_{t,\tau,\alpha} = \ell_{t,\tau,\alpha} - x \]

\[ m_{t,\tau,\alpha} = \log M_{t,\tau,\alpha}. \]

Here we are assuming that when the interval slightly changes from \((\tau', t')\) to \((\tau, t)\), \(\alpha'\) would not change much under the null hypothesis \((\alpha'\) is estimated from data stored in \((\tau', t')\)). Therefore, in the small neighborhood of \((\tau', t')\), we roughly regard \(\alpha\) as a constant. And we use the approximation:

\[ \sup_{t',\alpha'} e^{t',\alpha'} - \ell_{t,\tau,\alpha} \int_{t'} \int_{\alpha'} e^{t',\alpha'} - \ell_{t,\tau,\alpha} \, dt' \, d\alpha' \approx \sup_{t'} e^{t',\alpha'} - \ell_{t,\tau,\alpha} \int_{t'} e^{t',\alpha'} - \ell_{t,\tau,\alpha} \, dt' \]

The final representation of (32) will be consist of a large deviation exponential decay, given by \(e^{-x}\), and lower order contribution that reside in the expectation. The random variables in expectation are further dissected into random variables that are influenced mainly by local perturbations and the random variable that captures the main part of the variability. The latter is the random variable \(\bar{\ell}_{t,\tau,\alpha}\), and has expectation \((t - \tau)I - x\) under the alternative, and variance \((t - \tau)\sigma^2\). Here, we denote:

\[ I = \frac{E_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}]}{t - \tau}, \]

\[ \sigma^2 = \frac{\text{Var}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}]}{t - \tau}. \]

We call this random variable the global term. The other random variables are \(M_{t,\tau,\alpha}\) and \(S_{t,\tau,\alpha}\) and its log \(m_{t,\tau,\alpha}\). The local field \(\{\ell_{t',\alpha'} - \ell_{t,\tau,\alpha}\}\) are parameterized by \(t'\) when we fix \(t - \tau\).

We produce \(\hat{M}_{t,\tau,\alpha}\) and \(\hat{S}_{t,\tau,\alpha}\) by restricting the integral and maximization only to the range of parameter values that are at most \(\epsilon\) away from either \(\tau\) or \(t\). By localization theorem (Theorem 5.2 in [27]), under certain conditions, the local and global components are asymptotically independent, which informs:

\[ \mathbb{E}_{t,\tau,\alpha} \left[ \frac{M_{t,\tau,\alpha}}{S_{t,\tau,\alpha}} e^{-[\bar{\ell}_{t,\tau,\alpha} + m_{t,\tau,\alpha}]} : \bar{\ell}_{t,\tau,\alpha} + m_{t,\tau,\alpha} > 0 \right] \approx \mathbb{E}_{t,\tau,\alpha} \left[ \frac{M_{t,\tau,\alpha}}{S_{t,\tau,\alpha}} \right] \frac{1}{\sqrt{(t - \tau)\sigma^2}} \phi \left( \frac{(t - \tau)I - x}{\sqrt{(t - \tau)\sigma^2}} \right), \]

(40)

where we can further prove (see Appendix A.2) that the expected local rate \(\mathbb{E}_{t,\tau,\alpha} \left[ \frac{M_{t,\tau,\alpha}}{S_{t,\tau,\alpha}} \right]\) only depends on \(\alpha\) and is independent of \(t\):

\[ \mathbb{E}_{t,\tau,\alpha} \left[ \frac{M_{t,\tau,\alpha}}{S_{t,\tau,\alpha}} \right] = \nu \left( \frac{2\xi}{\eta^2} \right), \]

(41)
where

\[
\xi = -\frac{\mathbb{E}[\ell_{t,\tau,\alpha} - \mathbb{E}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}]]}{t - \tau}, \tag{42}
\]
\[
\eta^2 = \frac{\text{Var}[\ell_{t,\tau,\alpha}] + \text{Var}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}]}{t - \tau}. \tag{43}
\]

And the function \(\nu(\mu)\) is defined as:

\[
\nu(\mu) \approx \frac{(2/\mu)(\Phi(\mu/2) - 0.5)}{(\mu/2)\Phi(\mu/2) + \phi(\mu/2)}, \tag{44}
\]

where \(\Phi(x)\) is the cumulative distribution of the standard normal distribution and \(\phi(x)\) is the probability density function of standard normal distribution.

Then, we can write

\[
P\left[\mathop{\sup}_{t < m, \alpha \in \Theta} \ell_{t,\tau,\alpha} > x\right] \approx e^{-x} \int \int_{\alpha \in \Theta} \nu\left(\frac{2\xi}{\eta^2}\right) \frac{1}{\sqrt{(t - \tau)\sigma^2}} \phi\left(\frac{(t - \tau)I - x}{\sqrt{(t - \tau)\sigma^2}}\right) \, d\alpha \, dt \tag{45}
\]
\[
\approx me^{-x} \int_{\alpha \in \Theta} \nu\left(\frac{2\xi}{\eta^2}\right) \frac{1}{\sqrt{(t - \tau)\sigma^2}} \phi\left(\frac{(t - \tau)I - x}{\sqrt{(t - \tau)\sigma^2}}\right) \, d\alpha. \tag{46}
\]

Due to the fact that

\[
P\left[\mathop{\sup}_{t < m, \alpha \in \Theta} \ell_{t,\tau,\alpha} > x\right] = P[T < m], \tag{47}
\]

we can relate (45) to \(E[T]\). Note that we can write the tail probability (45) in a form \(P[T < m] = mC[1 + o(1)]\). We may see that \(T\) is asymptotically exponentially distributed and \(P[T < m] \to 1 - \exp(-Cm)\). As a result, \(E[T] \sim C^{-1}\). Overall, the expected stopping time would have the form:

\[
E[T] \approx e^x \left[\int_{\alpha \in \Theta} \nu\left(\frac{2\xi}{\eta^2}\right) \frac{1}{\sqrt{(t - \tau)\sigma^2}} \phi\left(\frac{(t - \tau)I - x}{\sqrt{(t - \tau)\sigma^2}}\right) \, d\alpha\right]^{-1}. \tag{48}
\]

Detailed computations regarding \(I, \sigma^2, \xi\) and \(\eta^2\) will be talked about in Appendix A.4.

A.2 Approximate local rate

To prove (41)-(44), we need to evaluate the mean and variance of the local field \(\{\ell_{t+\epsilon,\tau+\epsilon,\alpha} - \ell_{t,\tau,\alpha}\}\) after change-of-measures (i.e., under alternative distribution). Note that, from (6) we see the the log-likelihood ratio \(\ell_{t,\tau,\alpha}\) is an integration from time \(\tau\) to \(t\). Thus, we can dissect \(\ell_{t+\epsilon,\tau+\epsilon,\alpha}\) by dissecting the integration region in the following way

\[
\int_{\tau+\epsilon}^{t+\epsilon} = \int_{\tau+\epsilon}^{\tau+\epsilon^+} + \int_{\tau+\epsilon^+}^{t+\epsilon} + \int_{t+\epsilon^-}^{t+\epsilon}. \tag{49}
\]
where only the data stored in time interval \((\tau + \epsilon^+, t + \epsilon^-)\) are the common data which are also used to compute \(\ell_{t,\tau,0}\). Therefore,

\[
E_{t,\tau,0}[\ell_{t+\epsilon,\tau+\epsilon,a}] = E[\ell_{t+\epsilon,\tau+\epsilon,a}e^{\ell_{t,\tau,a}}] = E[\ell_{t+\epsilon,\tau+\epsilon,a} + \ell_{t,\tau,0}e^{\ell_{t,\tau,a}} + \ell_{t+\epsilon,t+\epsilon-,a}e^{\ell_{t,\tau,a}}] = E[\ell_{t+\epsilon,\tau+\epsilon,a}e^{\ell_{t,\tau,a}}] + \ell_{t,\tau,0}E[e^{\ell_{t,\tau,a}}].
\]

Note that if we define the null distribution density is \(f(X)\) and the alternative distribution density is \(g(X)\), obviously

\[
E[e^{\ell_{t,\tau,a}}] = \int \exp \left\{ \log \frac{g(X)}{f(X)} \right\} f(X)dX = \int \frac{g(X)}{f(X)}f(X)dX = \int g(X)dX = 1.
\]

Therefore, we have:

\[
E_{t,\tau,0}[\ell_{t+\epsilon,\tau+\epsilon,a}] = E[\ell_{t+\epsilon,\tau+\epsilon,a}] + \ell_{t,\tau,0}E[e^{\ell_{t,\tau,a}}] = -\epsilon^-E[\ell_{t,\tau,a}] + \epsilon^+\frac{E[\ell_{t,\tau,a}]}{t-\tau} + \epsilon^-\frac{E[\ell_{t,\tau,a}]}{t-\tau}.
\]

For the last equality, we are using the fact the both \(E[\ell_{t,\tau,a}]\) and \(E_{t,\tau,0}[\ell_{t,\tau,a}]\) are linear with time interval \((t-\tau)\), which will be proven in Appendix A.4. Finally we have:

\[
E_{t,\tau,0}[\ell_{t+\epsilon,\tau+\epsilon,a} - \ell_{t,\tau,a}] = (-\epsilon^- + \epsilon^+)\frac{E[\ell_{t,\tau,a}]}{t-\tau} - (\epsilon^+ - \epsilon^-)\frac{E[\ell_{t,\tau,a}]}{t-\tau} = \frac{E[\ell_{t,\tau,a}] - E_{t,\tau,0}[\ell_{t,\tau,a}]}{t-\tau}|\epsilon|.
\]

By Jensen’s inequality, we can prove that \(E[\ell_{t,\tau,a}] < 0\) and \(E_{t,\tau,0}[\ell_{t,\tau,a}] > 0\).

Similarly, we can derive the variance of the local field:

\[
\text{Var}_{t,\tau,0}[\ell_{t+\epsilon,\tau+\epsilon,a} - \ell_{t,\tau,a}] = \text{Var}_{t,\tau,0}[\ell_{t+\epsilon,\tau+\epsilon,a} - \ell_{t,\tau,a}] - \text{Var}_{t,\tau,0}[\ell_{t,\tau,a} - \ell_{t,\tau,a}] = \text{Var}[\ell_{t,\tau,a} - \ell_{t,\tau,a}] - \text{Var}[\ell_{t,\tau,a} - \ell_{t,\tau,a}] = \text{Var}[\ell_{t,\tau,a} - \ell_{t,\tau,a}] = (\epsilon^+ - \epsilon^-)\frac{\text{Var}[\ell_{t,\tau,a}]}{t-\tau} + (\epsilon^+ - \epsilon^-)\frac{\text{Var}[\ell_{t,\tau,a}]}{t-\tau} = \frac{\text{Var}[\ell_{t,\tau,a}]}{t-\tau} \left( \frac{\epsilon^+ - \epsilon^-}{\eta^2} \right).
\]

We are using the fact that both \(\text{Var}[\ell_{t,\tau,a}]\) and \(\text{Var}_{t,\tau,0}[\ell_{t,\tau,a}]\) are approximately linear with time interval \((t-\tau)\), which will be proven in Appendix A.4.

The above derivations show that the asymptotic distribution of \(\{\ell_{t+\epsilon,\tau+\epsilon,a} - \ell_{t,\tau,a}\}\), for small \(|\epsilon|\) is a two-sided brownian motion with a negative drift \(-\xi\). The variance of an increment of this brownian motion is \(\eta^2\). That is, the recentered process:

\[
\ell_{t+\epsilon,\tau+\epsilon,a} - \ell_{t,\tau,a} = B(\eta^2|\epsilon|) - \xi|\epsilon|
\]

with the equality meaning equality in distribution. According to Chapter 3 in [27], we can
approximate the local rate
\[
E_{t,\tau,\alpha} \left[ \frac{\mathcal{M}_{t,\tau,\alpha}}{S_{t,\tau,\alpha}} \right] = E_{t,\tau,\alpha} \left[ \sup_{\ell' \in [t,\tau,\alpha]} \frac{e^{\ell_{t,\tau,\alpha} - \ell_{t,\tau,\alpha}}}{\int_{\ell'} e^{\ell_{t,\tau,\alpha} - \ell_{t,\tau,\alpha}} d\ell'} \right] = \nu \left( \frac{\xi}{\eta^2} \right).
\] (65)

A.3 First- and second-order statistics for Hawkes processes

We need to characterize the first- and second-order statistics for Hawkes processes before evaluating \( I, \sigma^2, \xi \) and \( \eta^2 \), which will be used to compute expected stopping time (48). The following calculations mainly follow the work of [10]. [1] have the similar derivations.

For the defined one-dimensional Hawkes processes and high-dimensional Hawkes processes, if we choose kernel function \( \varphi(t) \) with \( \int \varphi(t) dt = 1 \), we will have the following two lemmas:

Lemma 2 (First-order statistics for Hawkes processes.). If the influence parameters satisfy \( \alpha \in (0,1) \) (one-dimension) or the spectral norm \( \rho(A) < 1 \) (high-dimension), then the Hawkes processes are asymptotically stationary and with stationary intensity \( m_t = E_{H_{t-}}[\lambda_t] \). We further have \( \bar{\lambda} := \lim_{t \to \infty} m_t = \frac{\mu}{1 - \alpha} \) (one-dimension) and \( \bar{\lambda} := \lim_{t \to \infty} m_t = (I - A)^{-1} \mu \) (high-dimension).

Lemma 3 (Second-order statistics for Hawkes processes.). For stationary Hawkes processes, the covariance intensity, which is defined as:
\[
c(t' - t) = Cov[\lambda_t, \lambda_{t'}] = Cov[dN_t, dN_{t'}] \frac{dt}{dt'}
\] will only depend on \( t' - t \). We can compute the covariance intensity for high-dimensional Hawkes processes as:
\[
c(\tau) = \begin{cases} 
\beta e^{-(I-A)\tau} A \left( I + \frac{1}{2} (I - A)^{-1} A \right) \text{diag}((I - A)^{-1} \mu), & \tau > 0; \\
\text{diag}((I - A)^{-1} \mu) \delta(\tau), & \tau = 0; \\
c(-\tau)^T, & \tau < 0.
\end{cases}
\] (67)

For one-dimensional Hawkes processes, we have:
\[
c(\tau) = \begin{cases} 
\frac{\alpha \beta (2 - \alpha) \mu}{2(1 - \alpha)^2} e^{-(1 - \alpha)\tau}, & \tau > 0; \\
\frac{\mu}{1 - \alpha} \delta(\tau), & \tau = 0; \\
c(-\tau), & \tau < 0.
\end{cases}
\] (68)

Next, we will prove Lemma 2 and 3.

Proof of Lemma 2. For the defined high dimensional Hawkes processes, by mean field approximation and define \( m_t = E_{H_{t-}}[\lambda_t] \), we have:
\[
m_t = \mu + A \int_{-\infty}^{t} \varphi(t-s) m_s ds
\] (69)
which can be written as
\[
m_t = \left( I + \sum_{n=1}^{\infty} A^n \int_{-\infty}^{t} \varphi^{(n)}(s) ds \right) \mu.
\] (70)
Define: $\Psi(t) = A\varphi(t) + A^2\varphi(t) \ast \varphi(t) + A^3\varphi(t) \ast \varphi(t) \ast \varphi(t) + \cdots = \sum_{n=1}^{\infty} A^n \varphi^{(n)}(t)$.

And we can write (70) as:

$$m_t = \left( I + \int_{-\infty}^{t} \Psi(s)ds \right) \mu.$$  \hfill (71)

Given a function $f(t)$, we denote its Laplace transform $\mathcal{L}(\cdot)$ as:

$$\tilde{f}(z) = \mathcal{L}(f(t)) = \int_{-\infty}^{\infty} f(t)e^{-zt}dt$$  \hfill (72)

Next, apply Laplace transform to both sides of equation (70). Clearly

$$\tilde{m}(z) = \frac{1}{z}(I - \frac{\beta}{z + \beta}A)^{-1}\lambda_0,$$  \hfill (73)

where

$$\Psi(z) = \sum_{n=1}^{\infty} \left( \frac{\beta}{z + \beta} \right)^n \cdot A^n = (I - \frac{\beta}{z + \beta}A)^{-1} - I.$$  \hfill (74)

By the property of Laplace transformation,

$$\tilde{\lambda} := \lim_{t \to \infty} m_t = \lim_{z \to 0} z\tilde{m}(z) = (I - A)^{-1}\mu.$$  \hfill (75)

For a special case where $d = 1$, we have $\tilde{\lambda} = \frac{\mu}{1-\alpha}$.

\textbf{Proof for Lemma 3.} For $\tau > 0$, we have:

$$c(\tau) = \frac{\mathbb{E}\left[ dN_{t+\tau}^TdN_t^T \right]}{(dt)^2} - \tilde{\lambda}\tilde{\lambda}^T$$  \hfill (76)

$$= \mathbb{E}\left[ \lambda_{t+\tau} \frac{dN_t^T}{dt} \right] - \tilde{\lambda}\tilde{\lambda}^T$$  \hfill (77)

$$= \mathbb{E}\left[ (\mu + A \int_{-\infty}^{t+\tau} \varphi(t + \tau - s)dN_s) \frac{dN_t^T}{dt} \right] - \tilde{\lambda}\tilde{\lambda}^T$$  \hfill (78)

$$= A \int_{-\infty}^{\tau} \varphi(\tau - s)c(s)ds$$  \hfill (79)

$$= A\varphi(\tau)\text{diag}(\tilde{\lambda}) + A \int_{-\infty}^{\tau} \varphi(\tau - s)c(s)ds$$  \hfill (80)

$$= A\varphi(\tau)\text{diag}(\tilde{\lambda}) + A \int_{0}^{\infty} \varphi(\tau + s)c(s)ds + A \int_{0}^{\tau} \varphi(\tau - s)c(s)ds.$$  \hfill (81)

For the last two equalities, we are using the relation, $c(-\tau) = c(\tau)^T$ and the fact that when $\tau = 0$, we have an extra term: $c(\tau) = \text{diag}(\tilde{\lambda})\delta(\tau)$.

Note that for Poisson processes, we will only have $c(\tau) = \text{diag}(\tilde{\lambda})\delta(\tau)$.

Plug in $\varphi(\tau) = \beta e^{-\beta\tau}$, we have:

$$c(\tau) = A\beta e^{-\beta\tau}\text{diag}(\tilde{\lambda}) + A \int_{0}^{\infty} \beta e^{-\beta(\tau + s)}c(s)ds + A \int_{0}^{\tau} \beta e^{-\beta(\tau - s)}c(s)ds.$$  \hfill (82)

Apply Laplace transform to both sides of (82), we have:

$$\tilde{c}(z) = \frac{\beta}{s + \beta}A\text{diag}(\tilde{\lambda}) + \frac{\beta}{s + \beta}A\tilde{c}(\beta) + \frac{\beta}{s + \beta}A\tilde{c}(z)$$  \hfill (83)
where
\[ \mathcal{L} \left( \int_0^\infty \beta e^{-\beta(\tau + s)} c(s) ds \right) = \mathcal{L} \left( \beta e^{-\beta \tau} \int_0^\infty e^{-\beta s} c(s) ds \right) = \mathcal{L} \left( \beta e^{-\beta \tau} \tilde{c}(\beta) \right) = \frac{\beta}{s + \beta} \tilde{c}(\beta). \] (84)

Replace \( z \) by \( \beta \), we can get the explicit expression for \( \tilde{c}(\beta) \):
\[ \tilde{c}(\beta) = \frac{1}{2} (I - A)^{-1} A \text{diag}(\bar{\lambda}). \] (85)

Therefore,
\[ \tilde{c}(z) = ((s + \beta) I - \beta A)^{-1} \beta A \left( I + \frac{1}{2} (I - A)^{-1} A \right) \text{diag}((I - A)^{-1} \mu) \] (86)

Next, perform reverse Laplace transform for \( \tilde{c}(z) \):
\[ c(\tau) = \mathcal{L}^{-1} (\tilde{c}(z)) = \beta e^{-\beta (I - A) \tau} A \left( I + \frac{1}{2} (I - A)^{-1} A \right) \text{diag}((I - A)^{-1} \mu), \quad \tau > 0. \] (87)

For a special case, when \( d = 1 \), we obtain:
\[ c(\tau) = \frac{\alpha \beta (2 - \alpha) \mu}{2(1 - \alpha)^2} e^{-\beta (1 - \alpha) \tau}, \quad \tau > 0. \] (88)

### A.4 Expectation and variance of log-likelihood ratio under null and alternative distributions

According to the definitions of \( I \), \( \sigma^2 \), \( \xi \) and \( \eta^2 \) \([38, 39, 42, 43]\), evidently, all the calculations will be finally boiled down to the evaluations of \( \mathbb{E}_{t, \tau, \alpha}[\ell_{t, \tau, \alpha}] \), \( \text{Var}_{t, \tau, \alpha}[\ell_{t, \tau, \alpha}] \), \( \mathbb{E}[\ell_{t, \tau, \alpha}] \) and \( \text{Var}[\ell_{t, \tau, \alpha}] \). That is, the expectation and variance of log-likelihood ratio under null and alternative distributions.

The main techniques will be used in evaluating these terms will be mean-field approximation, Delta method and the conclusions from Lemma 2 and 3. We constructed log-likelihood ratio based on two scenarios: data change from Poisson processes to Hawkes processes; data change from one Hawkes processes to another Hawkes processes with different influence parameters. The log-likelihood ratio established for these settings are slightly different, therefore we consider these cases one by one.

#### A.4.1 One-dimension: we want to detect changes from a one-dimensional Poisson process to a one-dimensional Hawkes process.

The hypothesis test problem is described as \([2]\), and the log-likelihood ratio is constructed as \([5]\).

Assuming stationary and \( t - \tau \) is large, we can approximate the stationary intensity for the alternative Hawkes processes to be \( \tilde{\lambda}^* \), which is defined as \( \tilde{\lambda}^* = \lim_{t \to \infty} m_t^* = \lim_{t \to \infty} \mathbb{E}_{H_t}[\lambda_t^*] \). We will use mean field approximation which assumes each stochastic process \( \lambda_t^* \) has small fluctuations around its mean \( \tilde{\lambda}^* \). That is the mean-field hypothesis \( \frac{\lambda_t^* - \tilde{\lambda}^*}{\tilde{\lambda}^*} \ll 1 \) holds. Then we compute the expectation of log-likelihood ratio under alterna-
tive distribution,

\[
E_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}] = E_{t,\tau,\alpha} \left[ \int_{\tau}^{t} \log (\lambda_s^*) dN_s - \int_{\tau}^{t} \log (\lambda_s) dN_s - \int_{\tau}^{t} (\lambda_s^* - \lambda_s) ds \right] \quad (89)
\]

\[
\approx E_{H_{t-}} \left[ \int_{\tau}^{t} \lambda_s^* \log (\lambda_s^*) ds - \int_{\tau}^{t} \lambda_s^* \log (\lambda_s) ds \right] - \int_{\tau}^{t} (m_s^* - \lambda_s) ds. \quad (90)
\]

Here, we are using the fact that under \(P_{t,\tau,\alpha}\), \(N(ds)\) is a Hawkes random field with conditional intensity \(\lambda_s^*\). From (89) to (90), more justifications can be found in [3, 4, 26]. In their work, they derived the average information gain per unit of time where the log-likelihood ratio is between a point process and a Poisson process.

Next, when \(t - \tau\) is large, we can approximate the stationary intensity for the alternative Hawkes processes to be \(\bar{\lambda}^*\). To approximate \(E_{H_{t-}} \left[ \int_{\tau}^{t} \lambda_s^* \log (\lambda_s^*) ds \right]\), we perform the first order taylor expansion for a new defined function \(f(\lambda_s) = \lambda_s^* \log (\lambda_s^*)\) around \(E_{H_{t-}}[\lambda_s^*] = \bar{\lambda}^*\) (Delta Method):

\[
\lambda_s^* \log (\lambda_s^*) \approx \bar{\lambda}^* \log (\bar{\lambda}^*) + [\log (\bar{\lambda}^*) + 1] (\lambda_s^* - \bar{\lambda}^*). \quad (91)
\]

Take expectation on both sides of the equation and use \(E_{H_{t-}}[\lambda_s^*] = \bar{\lambda}^*\),

\[
E_{H_{t-}} \left[ \int_{\tau}^{t} \lambda_s^* \log (\lambda_s^*) ds \right] \approx \int_{\tau}^{t} \bar{\lambda}^* \log (\bar{\lambda}^*) ds. \quad (92)
\]

Finally, we have:

\[
E_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}] \approx (t - \tau) \left[ \bar{\lambda}^* \log \left( \frac{\bar{\lambda}^*}{\mu} \right) - (\bar{\lambda}^* - \mu) \right] = (t - \tau) \left[ \frac{\mu}{1 - \alpha} \log \left( \frac{1}{1 - \alpha} \right) - \frac{\alpha}{1 - \alpha} \mu \right]. \quad (93)
\]

where \(I\) is the Kullback-Leibler information per time under the stationary assumption.

On the other hand, under the null distribution and given stationary assumption, we have:

\[
E[\ell_{t,\tau,\alpha}] = E \left[ \int_{\tau}^{t} \log \left( \frac{\lambda_s^*}{\lambda_s} \right) dN_s - \int_{\tau}^{t} (\lambda_s^* - \lambda_s) ds \right] \quad (94)
\]

\[
\approx E_{H_{t-}} \left[ \int_{\tau}^{t} \lambda_s \log \left( \frac{\lambda_s^*}{\lambda_s} \right) ds - \int_{\tau}^{t} (\lambda_s^* - \lambda_s) ds \right] \quad (95)
\]

\[
\approx (t - \tau) \left[ \mu \log \left( \frac{1}{1 - \alpha} \right) - \frac{\alpha}{1 - \alpha} \mu \right]. \quad (96)
\]

For the second equality we are using the fact that under \(P\), \(N(ds)\) is a Poisson random field with intensity \(\lambda_s\). For the last equality, we are using mean-field approximation.

Next, we compute the variance of log-likelihood ratio under null distribution and alternative distribution respectively. Under the alternative distribution,

\[
\int_{\tau}^{t} \log \left( \frac{\lambda_s^*}{\lambda_s} \right) dN_s - \int_{\tau}^{t} (\lambda_s^* - \lambda_s) ds \approx \int_{\tau}^{t} \left[ \lambda_s^* \log \left( \frac{\lambda_s^*}{\lambda_s} \right) - \lambda_s^* \right] ds + \lambda_s(t - \tau). \quad (97)
\]
Then the only random part is $f^t_\tau \left[ \lambda^*_s \log \left( \frac{\lambda^*_s}{\lambda^*_0} \right) - \lambda^*_s \right] ds$. Therefore,

$$\text{Var}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}] \approx \text{Var}_{H_{t,-}} \left[ \int_\tau^t \left[ \lambda^*_s \log \left( \frac{\lambda^*_s}{\lambda^*_0} \right) - \lambda^*_s \right] ds \right]. \quad (98)$$

Still using Delta method, we consider a new defined function with respect to $\lambda^*_s$

$$f(\lambda^*_s) = \lambda^*_s \log \left( \frac{\lambda^*_s}{\lambda^*_0} \right) - \lambda^*_s,$$  
and apply the first order taylor expansion around $E_{H_{t,-}}[\lambda^*_s] = \bar{\lambda}^*$:

$$f(\lambda^*_s) \approx f(\bar{\lambda}^*) + \left( \bar{\lambda}^* - \lambda^*_s \right) \log \left( \frac{\bar{\lambda}^*}{\lambda^*_0} \right). \quad (100)$$

From (100), we can get

$$\text{Var}_{H_{t,-}}[f(\lambda^*_s)] \approx \mathbb{E}_{H_{t,-}} \left[ (f(\lambda^*_s) - f(\bar{\lambda}^*))^2 \right] \approx \left[ \log \left( \frac{\bar{\lambda}^*}{\lambda^*_0} \right) \right]^2 \mathbb{E}_{H_{t,-}} \left[ (\lambda^*_s - \bar{\lambda}^*)^2 \right], \quad (101)$$

where $\mathbb{E}_{H_{t,-}}[ (\lambda^*_s - \bar{\lambda}^*)^2 ] = \text{Var}_{H_{t,-}}[\lambda^*_s]$. Note that the log-likelihood ratio is essentially an integration from $\tau$ to $t$. When computing the variance, we need to consider $\text{Cov}[\lambda^*_s, \lambda^*_{s+t}]$. Under the stationary assumption, from Lemma 3, we have expressions for $c(\tau) := \text{Cov}_{H_{t,-}}[\lambda^*_s, \lambda^*_{s+t}]$ which only depends on $\tau$. Therefore,

$$\text{Var}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}] \approx \left[ \log \left( \frac{\bar{\lambda}^*}{\lambda^*_0} \right) \right]^2 \int_\tau^t \int_\tau^t c(s' - s)dsds' \quad (102)$$

$$= \left[ \log \left( \frac{\bar{\lambda}^*}{\lambda^*_0} \right) \right]^2 \left[ \int_0^{t-\tau} \lambda^*_s ds + 2 \int_0^{t-\tau} \int_0^s c(v)dvds \right] \quad (103)$$

$$= (t - \tau)^2 \left[ \log \left( \frac{1}{1 - \alpha} \right) \right]^2 \left[ \frac{\mu}{1 - \alpha} + \frac{\alpha(2 - \alpha)\mu}{(1 - \alpha)^2} + \frac{\alpha(2 - \alpha)\mu e^{-\beta(1 - \alpha)(t - \tau)}}{\beta(1 - \alpha)^4(t - \tau)} \right]. \quad (104)$$

Moreover, $\alpha$ is usually a small number. When $(t - \tau)$ is a large number, ignore the small terms and further approximate:

$$\text{Var}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}] \approx (t - \tau)^2 \left[ \log \left( \frac{1}{1 - \alpha} \right) \right]^2 \left[ \frac{\mu}{1 - \alpha} + \frac{\alpha(2 - \alpha)\mu}{(1 - \alpha)^2} \right]. \quad (105)$$

On the other hand, under the null distribution, we have the variance of the log-likelihood ratio

$$\text{Var}[\ell_{t,\tau,\alpha}] \approx \left[ \log \left( \frac{\bar{\lambda}^*}{\lambda^*_0} \right) \right]^2 \int_\tau^t \lambda^*_s ds = (t - \tau)^2 \mu \left[ \log \left( \frac{1}{1 - \alpha} \right) \right]^2. \quad (107)$$
A.4.2 High-dimension: we want to detect changes from Poisson processes to Hawkes processes

The hypothesis test problem is described as (8), and the log-likelihood ratio is defined as (9). The derivations would follow the same strategy as the one-dimensional case, we just put the key results here.

For the expectation of the log-likelihood ratio under alternative distribution, we have:

\[
\mathbb{E}_{t,\tau,A}[\ell_{t,\tau,a}] \approx (t-\tau) \left[ \lambda^*\tau \left( \log(\lambda^*) - \log(\mu) \right) - e^\top(\lambda^* - \mu) \right] \tag{108}
\]

\[
= (t-\tau) \left[ (I-A)^{-1} \mu (\log((I-A)^{-1}\mu) - \log(\mu)) - e^\top ((I-A)^{-1}\mu - \mu) \right]. \tag{109}
\]

Under null, we have:

\[
\mathbb{E}[\ell_{t,\tau,a}] \approx (t-\tau) \left[ \mu^\top \left( \log(\lambda^*) - \log(\mu) \right) - e^\top(\lambda^* - \mu) \right] \tag{110}
\]

\[
= (t-\tau) \left[ \mu^\top \left( \log((I-A)^{-1}\mu) - \log(\mu) \right) - e^\top ((I-A)^{-1} - I)\mu \right]. \tag{111}
\]

For the variance of the log-likelihood ratio under alternative, we have

\[
\text{Var}_{t,\tau,A}[\ell_{t,\tau,a}] \tag{112}
\]

\[
= \text{Var}_{t,\tau,A} \left[ \int_{\tau}^{t} \log \left( \frac{\lambda_i(s)}{\mu_i} \right) dN^i_s \right] \tag{113}
\]

\[
= \sum_{i=1}^{d} \text{Var}_{t,\tau,A} \left[ \int_{\tau}^{t} \log \left( \frac{\lambda_i(s)}{\mu_i} \right) dN^i_s \right] + 2 \sum_{i<j} \text{Cov}_{t,\tau,A} \left[ \int_{\tau}^{t} \log \left( \frac{\lambda_i(s)}{\mu_i} \right) dN^i_s, \int_{\tau}^{t} \log \left( \frac{\lambda_j(s)}{\mu_j} \right) dN^j_s \right]. \tag{114}
\]

Note that from Lemma 3 for \( s > 0 \)

\[
c(s) = \beta e^{-\beta(I-A)s} A \left( I + \frac{1}{2} (I-A)^{-1} A \right) \text{diag} \left( (I-A)^{-1}\mu \right). \tag{115}
\]

To compute (114), we need to compute

\[
\int_{\tau}^{t} \int_{\tau}^{t} c(s' - s) ds ds' \tag{116}
\]

\[
= 2 \int_{0}^{t-\tau} \int_{0}^{s} c(v) dv ds \tag{117}
\]

\[
= 2\beta \int_{0}^{t-\tau} \int_{0}^{s} e^{-\beta(I-A)s} dv ds A \left( I + \frac{1}{2} (I-A)^{-1} A \right) \text{diag} \left( (I-A)^{-1}\mu \right) \tag{118}
\]

\[
= 2\beta \int_{0}^{t-\tau} \left( -\frac{1}{\beta} (I-A)^{-1} (e^{-\beta(I-A)s} - I) \right) ds A \left( I + \frac{1}{2} (I-A)^{-1} A \right) \text{diag} \left( (I-A)^{-1}\mu \right) \tag{119}
\]

\[
= 2 (I-A)^{-1} \int_{0}^{t-\tau} (I - e^{-\beta(I-A)s}) ds A \left( I + \frac{1}{2} (I-A)^{-1} A \right) \text{diag} \left( (I-A)^{-1}\mu \right) \tag{120}
\]

\[
\approx (t-\tau)(I-A)^{-1} A \left( 2I + (I-A)^{-1} A \right) \text{diag} \left( (I-A)^{-1}\mu \right). \tag{121}
\]
We have to note that when we consider \( \text{Cov}[dN_s^i, dN_s^j] \) (for the same dimension), we need to consider an extra term:
\[
\int_\tau^t \int_\tau^t \tilde{\lambda}^*(s' - s) ds' = \int_0^{t-\tau} \tilde{\lambda}^* ds = (t - \tau) \tilde{\lambda}^*. \tag{122}
\]
Arrange every terms, and use mean-field approximation and Delta method, we will finally have the following expressions:
\[
\text{Var}_{t, \tau, A} [\ell_{t, \tau, A}] \approx (t - \tau) \frac{\mathbf{e}^\top (\mathbf{H} \circ \mathbf{C}) \mathbf{e}}{\sigma^2}, \tag{123}
\]
where
\[
\mathbf{H} = \left[ \log \left( (I - \mathbf{A})^{-1} \mu \right) - \log(\mu) \right] \left[ \log \left( (I - \mathbf{A})^{-1} \mu \right) - \log(\mu) \right]^\top,
\]
\[
\mathbf{C} = (I - \mathbf{A})^{-1} \mathbf{A} \left( 2I + (I - \mathbf{A})^{-1} \mathbf{A} \right) \text{diag} \left( (I - \mathbf{A})^{-1} \mu \right) + \text{diag} \left( (I - \mathbf{A})^{-1} \mu \right). \tag{125}
\]
We compute the variance of the log-likelihood under null distribution. Note that when the data follow Poisson processes, we have \( \text{Cov}[N_i, N_j]_{i \neq j} = 0 \). Therefore,
\[
\text{Var}[\ell_{t, \tau, A}] \approx (t - \tau) \left[ \frac{\mathbf{\mu}^\top (\log(\tilde{\lambda}^*) - \log(\mu))^{(2)}}{\sigma^2} \right] \tag{126}
\approx (t - \tau) \left[ \frac{\mathbf{\mu}^\top (\log((I - \mathbf{A})^{-1} \mu) - \log(\mu))^{(2)}}{\sigma^2} \right], \tag{127}
\]
where the notation \( \mathbf{M}^{(2)} = \mathbf{M} \circ \mathbf{M} \) (Hadamard product).

**A.4.3 One-dimension:** we want to detect changes from a one-dimensional Hawkes process to another one-dimensional Hawkes process with different influence parameter.

Consider a hypothesis test as \((10)\) and the log-likelihood ratio is defined as \((11)\). Similarly, we compute the expectation of the log-likelihood ratio under alternative distribution
\[
\mathbb{E}_{t, \tau, \alpha} [\ell_{t, \tau, A}] = \mathbb{E}_{t, \tau, \alpha} \left[ \int_\tau^t \log(\tilde{\lambda}_s^*) dN_s - \int_\tau^t \log(\lambda_s) dN_s - \int_\tau^t (\lambda_s^* - \lambda_s) dS \right] \tag{128}
\approx \mathbb{E}_{\mathcal{H}_{1-\alpha}} \left[ \int_\tau^t \lambda_s^* \log(\lambda_s^*) ds - \int_\tau^t \lambda_s^* \log(\lambda_s) ds - \int_\tau^t (\lambda_s^* - \lambda_s) ds \right] \tag{129}
\approx (t - \tau) \left[ \frac{\lambda^* \log(\lambda^*) - \bar{\lambda} \log(\bar{\lambda}) - (\lambda^* - \bar{\lambda})}{\lambda^*} \right] \tag{130}
\approx (t - \tau) \left[ \frac{\mu}{1 - \alpha^*} \log \left( \frac{1 - \alpha}{1 - \alpha^*} \right) - \frac{\mu}{1 - \alpha^*} + \frac{\mu}{1 - \alpha} \right], \tag{131}
\]
where the first approximation is due to that under \( \mathbb{P}_{t, \tau, \alpha} \), \( N(ds) \) is a Hawkes random field with intensity \( \lambda_s^* \), and for the latter approximation, we are using mean field approximation and (multivariate) Delta Method given \( \mathbb{E}_{\mathcal{H}_{1-\alpha}} [\lambda^*(s)] = \bar{\lambda}^* \) and \( \mathbb{E}_{\mathcal{H}_{1-\alpha}} [\lambda_s] = \bar{\lambda} \). And for the stationary intensity, we have \( \bar{\lambda} = \frac{\mu}{1 - \alpha} \) and \( \bar{\lambda}^* = \frac{\mu}{1 - \alpha^*} \).
Next, we compute the expectation of the log-likelihood ratio under null distribution:

\[
\mathbb{E}[\ell_{t,\tau,\alpha}] = \mathbb{E}\left[ \int_{\tau}^{t} \log (\lambda^*_s \, dN_s) - \int_{\tau}^{t} \log (\lambda_s \, dN_s) - \int_{\tau}^{t} (\lambda^*_s - \lambda_s) \, ds \right] \tag{132}
\]

\[
\approx \mathbb{E}_{H^1_{\omega}}\left[ \int_{\tau}^{t} \lambda^*_s \log (\lambda^*_s) \, ds - \int_{\tau}^{t} \lambda_s \log (\lambda_s) \, ds - \int_{\tau}^{t} (\lambda^*_s - \lambda_s) \, ds \right] \tag{133}
\]

\[
\approx (t - \tau) \left[ \bar{\lambda} \log(\bar{\lambda}) - \bar{\lambda} \log(\bar{\lambda} - (\bar{\lambda}^* - \bar{\lambda})) \right] \tag{134}
\]

\[
= (t - \tau) \left\{ \frac{\mu}{1 - \alpha} \log \left( \frac{1 - \alpha}{1 - \alpha^*} \right) - \frac{\mu}{1 - \alpha^*} + \frac{\mu}{1 - \alpha} \right\}. \tag{135}
\]

We compute the variance of the log-likelihood ratio under alternative distribution. Under alternative distribution

\[
\ell_{t,\tau,\alpha} = \int_{\tau}^{t} \log (\lambda^*_s \, dN_s) - \int_{\tau}^{t} \log (\lambda_s \, dN_s) - \int_{\tau}^{t} (\lambda^*_s - \lambda_s) \, ds \tag{136}
\]

\[
\approx \int_{\tau}^{t} \left[ \lambda^*_s \log (\lambda^*_s) - \lambda^*_s \log (\lambda_s) - \lambda^*_s + \lambda_s \right] ds. \tag{137}
\]

Next, we perform the first order taylor expansion to the new defined multivariate function with respect to \( \lambda^*_s \) and \( \lambda_s \):

\[
f(\lambda^*_s, \lambda_s) \approx f(\bar{\lambda}^*, \bar{\lambda}) + \log \left( \frac{\bar{\lambda}^*}{\lambda} \right) (\lambda^*_s - \bar{\lambda}^*) + \left( 1 - \frac{\bar{\lambda}^*}{\lambda} \right) (\lambda_s - \bar{\lambda}). \tag{138}
\]

Based on that, we will have

\[
\text{Var} [f(\lambda^*_s, \lambda_s)] = \mathbb{E} \left[ (f(\lambda^*_s, \lambda_s) - f(\bar{\lambda}^*, \bar{\lambda}))^2 \right] \tag{139}
\]

\[
\approx \left[ \log \left( \frac{\bar{\lambda}^*}{\lambda} \right) \right]^2 \text{Var}[\lambda^*_s] + \left( 1 - \frac{\bar{\lambda}^*}{\lambda} \right)^2 \text{Var}[\lambda_s]. \tag{140}
\]

Note that the null intensity \( \lambda_s \) is independent of the alternative intensity \( \lambda^*_s \). Finally, we have:

\[
\text{Var}_{t,\tau,\alpha}[\ell_{t,\tau,\alpha}] \tag{141}
\]

\[
\approx \text{Var}_{t,\tau,\alpha}\left[ \int_{\tau}^{t} \lambda^*_s \log (\lambda^*_s) \, ds - \int_{\tau}^{t} \lambda^*_s \log (\lambda_s) \, ds - \int_{\tau}^{t} (\lambda^*_s - \lambda_s) \, ds \right] \tag{142}
\]

\[
\approx \left[ \log \left( \frac{\bar{\lambda}^*}{\lambda} \right) \right]^2 \int_{\tau}^{t} \int_{\tau}^{t} c'(s' - s) \, ds \, ds' + \left( 1 - \frac{\bar{\lambda}^*}{\lambda} \right)^2 \int_{\tau}^{t} \int_{\tau}^{t} c(s' - s) \, ds \, ds' \tag{143}
\]

\[
\approx (t - \tau) \left[ \log \left( \frac{1 - \alpha}{1 - \alpha^*} \right) \right]^2 \left[ \frac{\mu}{1 - \alpha^*} + \frac{\alpha^*(2 - \alpha^*) \mu}{(1 - \alpha^*)^3} \right] + \left( 1 - \frac{1 - \alpha}{1 - \alpha^*} \right)^2 \left[ \frac{\mu}{1 - \alpha} + \frac{\alpha(2 - \alpha) \mu}{(1 - \alpha)^3} \right]. \tag{144}
\]

Just like previously, we’ve ignored some small terms.

Similarly, we can compute the variance of the log-likelihood ratio under null distribution.
Under null distribution,
\[
\ell_{t,\tau,\alpha} \approx \int_{\tau}^{t} \lambda_s \log(\lambda_s^*) - \lambda_s \log(\lambda_s) - \lambda_s^* + \lambda_s \, ds. 
\tag{145}
\]

Still perform the first order Taylor expansion to the new defined function:
\[
f(\lambda_s^*, \lambda_s) \approx f(\bar{\lambda}^*, \bar{\lambda}) + \left( \frac{\bar{\lambda}}{\lambda^*} - 1 \right) (\lambda_s^* - \bar{\lambda}^*) + \log \left( \frac{\bar{\lambda}}{\lambda} \right) (\lambda_s - \bar{\lambda}). 
\tag{146}
\]

Therefore, using multivariate Delta method,
\[
\text{Var}[f(\lambda_s^*, \lambda_s)] = E \left[ (f(\lambda_s^*, \lambda_s) - f(\bar{\lambda}^*, \bar{\lambda}))^2 \right] \approx \left( \frac{\bar{\lambda}}{\lambda} - 1 \right)^2 \text{Var}[\lambda_s^*] + \left( \log \left( \frac{\bar{\lambda}}{\lambda} \right) \right)^2 \text{Var}[\lambda_s]. 
\tag{147}
\]

Finally we will obtain
\[
\text{Var}[\ell_{t,\tau,\alpha}] \approx \left( \frac{\bar{\lambda}}{\lambda^*} - 1 \right)^2 \int_{\tau}^{t} \int_{\tau}^{t} c(s' - s) ds'ds' + \left[ \log \left( \frac{\bar{\lambda}}{\lambda} \right) \right]^2 \int_{\tau}^{t} \int_{\tau}^{t} c(s' - s) ds'ds' 
\approx (t - \tau) \left[ 1 - \frac{1 - \alpha^*}{1 - \alpha} \right]^2 \left[ \frac{\mu}{1 - \alpha^*} + \frac{\alpha^*(2 - \alpha^*)\mu}{(1 - \alpha^*)^3} \right] + \left[ \log \left( \frac{1 - \alpha}{1 - \alpha^*} \right) \right]^2 \left[ \frac{\mu}{1 - \alpha} + \frac{\alpha(2 - \alpha)\mu}{(1 - \alpha)^3} \right]. 
\tag{149}
\]

\[
\sigma^2 = e^\top (G \circ C^* + F \circ C) e, 
\tag{155}
\]
\[
\sigma_0^2 = e^\top (R \circ C^* + G \circ C) e, 
\tag{156}
\]

A.4.4 High-dimension: we want to detect the changes from one multivariate Hawkes process to another multivariate Hawkes process with different influence matrix.

The hypothesis test is constructed as \[12\] and the log-likelihood ratio is defined as \[13\]. The main computation results are:
\[
I = \dot{\lambda}^* \top \left[ \log \dot{\lambda}^* - \log \bar{\lambda} \right] - e^\top [\dot{\lambda}^* - \bar{\lambda}], 
\tag{151}
\]
\[
I_0 = \dot{\lambda} \top \left[ \log \dot{\lambda} - \log \bar{\lambda} \right] - e^\top [\dot{\lambda} - \bar{\lambda}], 
\tag{152}
\]
where
\[
\dot{\lambda}^* = (I - A^*)^{-1} \mu, 
\tag{153}
\]
\[
\dot{\lambda} = (I - A)^{-1} \mu. 
\tag{154}
\]

And
\[
\sigma^2 = e^\top (G \circ C^* + F \circ C) e, 
\tag{155}
\]
\[
\sigma_0^2 = e^\top (R \circ C^* + G \circ C) e, 
\tag{156}
\]
where

\[ G_{ij} = \log \left( \frac{\lambda_i^*}{\lambda_i} \right) \log \left( \frac{\lambda_j^*}{\lambda_j} \right), \quad 1 \leq i \leq j \leq d, \]  

(157)

\[ F_{ij} = \left( 1 - \frac{\lambda_i^*}{\lambda_i} \right) \left( 1 - \frac{\lambda_j^*}{\lambda_j} \right), \quad 1 \leq i \leq j \leq d, \]  

(158)

\[ R_{ij} = \left( \frac{\lambda_i}{\lambda_i^*} - 1 \right) \left( \frac{\lambda_j}{\lambda_j^*} - 1 \right), \quad 1 \leq i \leq j \leq d. \]  

(159)

and

\[ C^* = (I - A^*)^{-1} A^* \left( 2I + (I - A^*)^{-1} A^* \right) \text{diag} \left( (I - A^*)^{-1} \mu \right) + \text{diag} \left( (I - A^*)^{-1} \mu \right), \]  

(160)

\[ C = (I - A)^{-1} A \left( 2I + (I - A)^{-1} A \right) \text{diag} \left( (I - A)^{-1} \mu \right) + \text{diag} \left( (I - A)^{-1} \mu \right). \]  

(161)

In the end, according to the definitions of \( \xi \) and \( \eta \), we have the following relation,

\[ \xi = I - I_{H_0}, \]  

(162)

\[ \eta = \sigma^2 + \sigma^2_{H_0}, \]  

(163)

from which we can obtain \( \xi \) and \( \eta \).