Fast Estimation of Multivariate Spatiotemporal Hawkes Processes and Network Reconstruction

Baichuan Yuan, Frederic P. Schoenberg , Andrea L. Bertozzi
University of California, Los Angeles

March 15, 2019

Abstract

We present a fast and accurate estimation method for multivariate Hawkes processes, a type of self-exciting point process that is widely used in seismology, criminology, finance, and many other areas. There are two major ingredients. The first is an analytic derivation of the likelihood-based estimation, which directly computes exact maximum likelihood estimates of the nonparametric triggering density. We develop it for the multivariate case and add regularization to improve stability and robustness. The second is a moment-based method for background rate and triggering matrix estimation, which is extended here for the spatiotemporal case. Our method combines them together in an efficient way and we prove the consistency of this new approach. Extensive numerical experiments, with synthetic data and real-world social network data, show that our method substantially improves the accuracy, scalability and computational efficiency of prevailing estimation approaches. Moreover, it greatly boosts the performance of Hawkes process-based models on social network reconstruction and helps to understand the spatiotemporal triggering dynamics over social media.

Keywords: non-parametric estimation, L_2 regularization, point processes, social network, cumulants

1 Introduction

The Spatiotemporal Hawkes (ST-Hawkes) process has been widely used to model and forecast clustered point process data in the study of earthquakes [32], crimes [28], invasive species [5], terrorist attacks [34], infectious disease [38], and finance [3]. These models, which are characterized by a *triggering density* describing how the occurrence of one event may spark future events nearby, have contributed to the rise of predictive policing [33], resulting in real-world impacts on the crime rate in Los Angeles [29]. Recently, multivariate Hawkes processes, which can incorporate accompanying information on each event such as the type of crime or magnitude of an earthquake, have been the subject of significant research in the areas of criminology [27], finance [3], neuroscience [8], and text analysis [12]. Applications include network reconstruction [24, 16, 18, 42, 25], causal inference [1, 14, 6] and social media cascade modeling [21, 15].

Much of this recent research has been fueled by advances in the nonparametric estimation of Hawkes processes, and in particular by the landmark work of Marsan and Lengliné [26], who detailed a method for estimating the triggering in a ST-Hawkes process by assuming the triggering density to be a step function and then estimating the step heights via maximum likelihood estimation (MLE). Such nonparametric estimation methods allow the triggering density to be estimated without assuming a particular parametric form which may be subject to misspecification or over-fitting, which can be very serious problems, especially in social science applications [42]. Instead, the data drive the estimation of the triggering density, and this is especially attractive for use with the large data sets that are increasingly becoming available in applications. Unfortunately, however, a major limitation of current nonparametric estimation methods is their computational complexity and lack of speed, as existing methods are mainly based on maximum likelihood estimation (MLE) [35], or variants such as EM-type algorithms [41, 26], which are typically non-convex problems without closed form solutions. For applications to crimes or to social media, for instance, catalogs of millions of ST events are often the subject of study, and each calculation of the likelihood function with N events requires at least $O(N^2)$ time. In such situations, the estimation of the triggering density using existing methods can be infeasible. As a result, it is important to develop better alternatives to current MLE-based methods [38].

Recent developments in the nonparametric estimation of the Hawkes process provide new insights for this problem, including an analytic method for computing the MLE of the triggering density in the special case where the adjacency matrix is invertible [40], and generalized moments methods (GMM) for the estimation of the triggering matrix [1]. In this paper, we propose a new, highly computationally efficient, scalable nonparametric estimator for ST-Hawkes processes, based on a blend of these recent ideas with modern advances in the regularization and inversion of sparse matrices.

The structure of this paper is as follows. We first review background material in Section 2. In Section 3.1, we extend the analytic formula for the MLE of the step heights in the triggering density [40] to the multivariate ST case and in Section 3.2 we greatly improve the stability of the resulting estimator using regularization. We next extend the cumulant-based estimators of [1] to the multivariate ST case and derive GMM estimators of the triggering matrix in this context in Section 3.3. We then combine the MLE estimators with GMM estimators to obtain a scalable, computationally efficient estimator, and we prove its consistency under general conditions in Section 3.4. The computational complexity of our approach is analyzed in Section 3.5. The performance of this estimator is inspected using a variety of synthetic and real social network datasets in Section 4, where we show that the proposed estimator has a computation complexity linear in the number of events N, allowing one to explore applications to large data sets with millions of events, and is shown to outperform current state-of-the-art methods in terms of both accuracy in network reconstruction and computation time. Finally, we conclude and discuss important directions for future research in Section 5.

2 Background on Multivariate Hawkes Processes and their Nonparametric Estimation

In this section, we review multivariate Hawkes processes and review previous research on their estimation methods, focusing especially on MLE and GMM.

A point process [11, 10] is a σ-finite collection of points $\{\tau_1, \tau_2, ...\}$ occurring in some metric space. While the definitions and results below can be extended quite readily to

other spaces, we will assume for simplicity throughout that the metric space is a bounded interval $[0, T]$ in time or a bounded interval $B \times [0, T]$ in space-time. A temporal or ST point process is typically modeled via its conditional intensity, $\lambda(t)$ or $\lambda(s, t)$, which represents the infinitesimal rate at which points are accumulating at the particular location in time or space-time, given information on all points occurring prior to time t. Simple point processes are uniquely characterized by their conditional intensity [11]; for models for non-simple point processes, see [36].

Hawkes processes are typically characterized via their conditional intensities. We refer readers to [11, 7] for details about these concepts. For a simple temporal Hawkes process [19], the conditional intensity of events at time t can be written

$$\lambda(t) = \mu + K \int_0^t g(t - t')dN(t'), \tag{1}$$

where $\mu > 0$, is the background rate, $g(v) \geq 0$ is the *triggering density* satisfying $\int_0^\infty g(v)dv = 1$ which describes the conductivity of events, and the constant K is the productivity, which is typically required to satisfy $0 \leq K < 1$ in order to ensure stationarity and subcriticality [19].

A *multivariate* temporal Hawkes process is conveniently viewed as a sequence of temporal point processes indexed by $u = 1, ..., U$, where each subprocess N_u has conditional intensity

$$\lambda_u(t) = \mu_u + \sum_{t_k < t} K_{u_k,u}\, g_{u_k}(t - t_k), \tag{2}$$

and the N points of the entire process may conveniently be labelled (t_k, u_k), for $k = 1, ..., N$, where t_k indicates the time of point k, and u_k indicates the index dictating to which subprocess the point belongs. The idea behind equation (2) is that the triggering density g_{u_k} and productivity $K_{u_k,u}$ may depend on the index of the point t_k.

In the model (2), μ_u is the background rate, indicating the rate at which points of mark u occur, absent any other prior events. For simplicity, one traditionally assumes a uniform background rate in time. $\boldsymbol{K} \in \mathbb{R}^{U \times U}$ is the triggering matrix, where $K_{u,v}$ is the expected number of events of index v that are triggered by one event of index u. This

triggering effect, in this temporal-only case, is closely related to Granger causality [17]. In fact, subprocess u does not Granger-cause subprocess v if and only if $K_{u,v} = 0$ [14]. Similarly, for stationarity and subcriticality, $\boldsymbol{K}$ needs to satisfy $\|\boldsymbol{K}\| < 1$, where $\|\boldsymbol{K}\|$ is the spectral norm of $\boldsymbol{K}$.

In nonparametric estimation of g, one typically assumes that each subprocess has the same piecewise-constant triggering densities $g_{u_k}(t) = g(t)$ which control how quickly the rate $\lambda_u(t)$ returns to its baseline level μ_u after an event occurs. One can estimate the parameters $\boldsymbol{\mu} = (\mu_u)_u$, $\boldsymbol{K}$, and the triggering densities g via MLE [31] or minimize a regression loss [8]. Here we focus on the MLE approach. The log-likelihood function of the intensity function (2) becomes

$$l = \sum_{k=1}^{N} log(\lambda_{u_k}(t_k)) - \sum_{u=1}^{U} \int_0^T \lambda_u dt. \tag{3}$$

One can directly maximize this function using off-the-shelf optimization methods or the EM-type algorithm proposed in [41]. See [42] for details about the derivation of the EM-type algorithm for ST-Hawkes processes. Another MLE-based approach, based on their analytic derivation of MLE, is first proposed in [40] for the univariate case ($U = 1$). They found that one can solve the MLE problem via solving linear equations in g and two additional linear equations for the background rate μ and productivity K. However, for the multivariate case, the coefficients of these equations depend on the triggering matrix $\boldsymbol{K}$ and it is no longer a linear system. Also, there is the problem of stability when the matrix of the linear system is singular or nearly singular. The inversion of the matrix is a major problem [40] in its implementation in practice, and in Section 3.2 we present the solution to this problem via regularization.

Another kind of estimation method [1, 4] is based on GMM using *cumulants* of Hawkes processes. Define $\mathbf{R} = (\mathbf{I} - \mathbf{K}^T)^{-1}$, where $\mathbf{I}$ is the identity matrix. As an alternative to the moments, the first, second and third *cumulant* of Hawkes process $\boldsymbol{\Lambda}$, $\boldsymbol{C}$ and $\boldsymbol{\Gamma}$ have the following relationships [1] with $\boldsymbol{R}$

$$\boldsymbol{\Lambda}(i) = \Lambda^i = \sum_{m=1}^{U} R^{im}\mu_m, \tag{4}$$

$$\boldsymbol{C}(i,j) = C^{ij} = \sum_{m=1}^{d} \Lambda^m R^{im} R^{jm}, \tag{5}$$

$$\boldsymbol{\Gamma}(i,j,k) = \Gamma^{ijk} = \sum_{m=1}^{d} (R^{im} R^{jm} C^{km} + R^{im} C^{jm} R^{km} + C^{im} R^{jm} R^{km} - 2\Lambda^m R^{im} R^{jm} R^{km}). \tag{6}$$

Here $R^{im} = \boldsymbol{R}(i,m)$. Although the definition and numerical estimations of the cumulants are different for the ST case, the above formulas still hold because the spatial information can be viewed as "marks" of the temporal point process.

The idea of GMM is to estimate the cumulants numerically from the data and then obtain the triggering matrix $\boldsymbol{K}^T = \mathbf{I} - \mathbf{R}^{-1}$ by minimizing the approximation error of the cumulants with some scaling coefficient κ

$$L(\boldsymbol{R}) = (1-\kappa)\|\boldsymbol{R}^{\odot 2}\hat{\boldsymbol{C}}^T + 2(\boldsymbol{R} \odot (\hat{\boldsymbol{C}} - \boldsymbol{R}\hat{\boldsymbol{L}}))\boldsymbol{R}^T - \hat{\boldsymbol{\Gamma}^{\boldsymbol{c}}}\|_2^2 + \kappa\|\boldsymbol{R}\hat{\boldsymbol{L}}\boldsymbol{R}^T - \hat{\boldsymbol{C}}\|_2^2. \tag{7}$$

Here $\odot$ is the Hadamard product and $\hat{\boldsymbol{\Gamma}^{\boldsymbol{c}}} = \hat{\boldsymbol{\Gamma}}(i,i,k)$. Given the estimated $\tilde{\boldsymbol{R}}$, we also have $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{\Lambda}}$ from the cumulants equation (4). This provides a fast estimation procedure for both $\boldsymbol{\mu}$ and $\boldsymbol{K}$. But it does not estimate the triggering density, which plays an important role in the dynamics of the point process. In applications such as stochastic declustering [43], it is necessary to estimate triggering densities from the data. Some other moment-based methods [4] can estimate both of them at the cost of high computation time.

3 Proposed Methods for Multivariate ST-Hawkes

In this section, we extend the previous discussion to the case of *multivariate* ST-Hawkes processes and derive a fast estimation method via extending and combining the two approach (MLE and GMM) discussed above. We recommend interested readers to check [35, 39] which provide comprehensive reviews of ST point processes. The focus of our method is to reduce the computational burden of the inference as well as improve the model estimation accuracy. Our motivation is from the application of network reconstruction. Previous studies have shown the ability of Hawkes process models to uncover the underlying connections

between nodes (such as social media users [42], neurons [8], email users [16] and crime [24]). It is essential to develop a scalable method because one often encounters data sets with thousands of nodes (large U) and millions of associated ST events (very large N).

We consider a simple multivariate ST-Hawkes process with a spatially isotropic triggering density $g(x, y, t)$ – i.e., $g(x, y, t) = g(r, t), r = \sqrt{x^2 + y^2}$ (g is only a function of time and distance). For each subprocess $u = 1, ..., U$, the conditional intensity characterizing the multivariate ST-Hawkes process is assumed to have the form

$$\lambda_u(x, y) = \mu_u(x, y) + \sum_{t_k<t} K_{u_k u} g(d_k, t - t_k) \tag{8}$$

where (t_k, x_k, y_k, u_k), for $k = 1, ..., N$, denote the N observed events in $B \times [0, T]$ and $d_k = \sqrt{(x_k - x)^2 + (y_k - y)^2}$. Current MLE-based methods such as the EM-type algorithm [41, 42] are not well-suited for large-scale problems due to its $O(N^3)$ computational complexity [1]. Also, in many applications, it is difficult to determine the appropriate triggering density $g(r, t)$. Our proposed method has a linear $O(N)$ complexity and learns triggering densities directly from data. Specifically, we estimate $g(r, t)$ nonparametrically from MLE and $\boldsymbol{K}$, $\boldsymbol{\mu}$ from GMM. This combined method gives a fast and complete estimation of the ST-Hawkes process.

3.1 ST Triggering Density Estimation

We extend the analytic method, first proposed in [40] for the univariate temporal case, to the case of multivariate ST Hawkes processes.

First, we review the derivation of analytic estimates of the triggering function for the multivariate temporal Hawkes process (2). Assuming that $\boldsymbol{\mu}$ and $\mathbf{K}$ are given or well-estimated by other means, assume the only variables here to be estimated are the heights of the step function comprising the triggering density $g(t) = \sum_{m=1}^{N_t} g_m \mathbb{1}_{t\in(\tau_m,\tau_{m+1})}$ with N_t grids $U_m = \{t \mid t \in (\tau_m, \tau_{m+1})\}, m = 1, ..., N_t$ in time. One seeks to obtain the step heights of the triggering density via maximizing the log-likelihood function. The log-likelihood function (from (3))

$$l = \sum_{k=1}^{N} log(\lambda_{u_k}(t_k)) - \sum_{u=1}^{U} (\mu_u T + \sum_{m=1}^{N_t} g_m \delta_m \sum_{k=1}^{N} K_{u_k u}) \tag{9}$$

is concave w.r.t $\{g_m\}_m$, and we take the derivative w.r.t. g_m

$$0 = \frac{\partial l}{\partial g_m} = \sum_{(t_j - t_i) \in U_m} \frac{K_{u_i u_j}}{\lambda_{u_j}(t_j)} - \sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m, \tag{10}$$

where $\delta_m = \tau_{m+1} - \tau_m$. Using the notation $\boldsymbol{\lambda} = \{\lambda_{u_j}(t_j)\}_j$, $\boldsymbol{A}(k,j) = \sum_{t_j - t_i \in U_k} K_{u_i u_j}$, $\boldsymbol{\beta} = \{g_m\}_m$ and $\boldsymbol{b} = \{\sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m\}_m$, we obtain a matrix representation of equation (10)

$$0 = \boldsymbol{A}(1/\boldsymbol{\lambda}) - b. \tag{11}$$

Here $1/\boldsymbol{\lambda}$ is the element-wise reciprocal. The solution of (11) yields an estimate of $\boldsymbol{\lambda}$. Further, equation (2) can be rewritten as

$$\boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{A}^T \boldsymbol{\beta}. \tag{12}$$

Solving this equation using the estimate of $\boldsymbol{\lambda}$ from (11) provides the maximum likelihood estimate of $\boldsymbol{\beta}$.

We now focus on the *multivariate* ST-Hawkes process with a piecewise-constant ST triggering density $g(r,t)$. We simply assume a uniform background rate $\mu_u(x,y) = \mu_u$. For each subprocess $u = 1, ..., U$, the conditional intensity satisfies

$$\lambda_u(x,y,t) = \mu_u + \sum_{t_k < t} K_{u_k u} \sum_{m=1}^{N_t} \sum_{n=1}^{N_r} g_{mn} \mathbb{1}_{t_k - t \in (\tau_m, \tau_{m+1})} \mathbb{1}_{d_k \in (r_n, r_{n+1})}. \tag{13}$$

Here $d_k = \sqrt{(x_k - x)^2 + (y_k - y)^2}$ and g is defined on a 2-D $N_r \times N_t$ grids with $V_n = \{d_k \mid d_k \in (r_n, r_{n+1})\}, n = 1, ..., N_r$ in distance and $U_m = \{t_k - t \mid t_k - t \in (\tau_m, \tau_{m+1})\}, m = 1, ..., N_t$ in time. The log-likelihood function of this intensity function is [37]

$$\begin{aligned} l &= \sum_{k=1}^{N} log(\lambda_{u_k}(x_k, y_k, t_k)) - \sum_{u=1}^{U} \iint_B \int_0^T \lambda_u(x,y,t) dt dx dy. \\ &= \sum_{k=1}^{N} log(\lambda_{u_k}(x_k, y_k, t_k)) - \sum_{u=1}^{U} (\mu_u |B| T + \sum_m \sum_n g_{mn} \delta_m \triangle_n \sum_{k=1}^{N} K_{u_k u}) \end{aligned} \tag{14}$$

where $|B|$ is the area of B, $\delta_m = \tau_{m+1} - \tau_m$ and $\triangle_n = \pi(r_{n+1}^2 - r_n^2)$.

Assuming that $\boldsymbol{\mu}$ and $\boldsymbol{K}$ are given, the only variable here are $\{g_{mn}\}_{m,n}$. Maximizing the log likelihood function will give us the estimation of the triggering density g. Since (14) is concave, we take the derivative of equation w.r.t g_{mn}

$$0 = \frac{\partial l}{\partial g_{mn}} = \sum_{(t_j - t_i) \in U_m, d_{ij} \in V_n} \frac{K_{u_i u_j}}{\lambda_{u_j}(x_j, y_j, t_j)} - \sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m \triangle_n, \tag{15}$$

with $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. In the same manner, we define $\boldsymbol{\lambda} = \{\lambda_{u_j}(x_j, y_j, t_j)\}_j$, $\boldsymbol{A}(k(m,n), j) = \sum_{t_j - t_i \in U_m, d_{ij} \in V_n} K_{u_i u_j}$, $\boldsymbol{\beta} = (g_{mn})_{k(m,n)}$ and $\boldsymbol{b} = (\sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m \triangle_n)_{k(m,n)}$ with the index $k(m,n) = N_r(m-1)+n$. Then we obtain the matrix representation of (15) and (13) as

$$0 = \boldsymbol{A}(1/\boldsymbol{\lambda}) - \boldsymbol{b}. \tag{16}$$

$$\boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{\beta}. \tag{17}$$

Finally we can estimate $\boldsymbol{\beta}$ via solving the above linear equations separately.

3.2 Regularization for Linear System

As noted in [40], in many applications, the matrix $\boldsymbol{A}$ in (16) and (17) is often ill-conditioned or singular, even with a careful selection of the 2-D grids U_m and V_n . Further, even when it can be obtained, the direct inverse $\boldsymbol{A}^{-1}\boldsymbol{b}$ (or pseudo inverse $(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}\boldsymbol{b}$) can give unstable results due to overfitting. In order to solve the linear equations in a stable and robust fashion, we use regularization procedures to find meaningful approximate solutions.

More specifically, we propose the use of the Tikhonov regularization method [30] with its analytic solution. For example, with the regularization, solving (16) becomes this minimization problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \|\boldsymbol{\Gamma}\boldsymbol{x}\|^2 \tag{18}$$

for Tikhonov matrix $\boldsymbol{\Gamma} = \alpha\boldsymbol{I}$. This is essentially an L_2 regularization, giving preference to solutions with smaller norms. L_1 regularization will typically give a sparse solution with many zero entities. It does not work here due the fact that each element in $\boldsymbol{x} = 1/\boldsymbol{\lambda}$ is positive and nonzero. Further one could use other Tikhonov matrices to guarantee smoothness if the underlying vector is believed to be mostly continuous. Instead of this, for the estimation of the triggering density, we smooth g with the following post-processing approach.

In addition, we assume that the triggering density is separable in space and time [32]. As a result, we can decompose the triggering density $g(r,t)$ into the spatial triggering density $f(r)$ and temporal triggering density $h(t)$ – i.e., $g(r,t) = f(r)h(t)$. If we reshape the N_rN_t-by-1 vector $\boldsymbol{\beta}$ as a N_r-by-N_t matrix $\boldsymbol{B}$, then estimating the spatial and temporal

triggering density becomes the following unmixing problem

$$\min_{\boldsymbol{f}\geq 0,\boldsymbol{h}\geq 0} \|\boldsymbol{B}-\boldsymbol{f}\boldsymbol{h}\|^2. \tag{19}$$

Here $\boldsymbol{B}$ is a nonnegative matrix based on the definition of $g(r,t)$ (triggering density function), $\boldsymbol{f}$ is a nonnegative N_r-by-1 vector and $\boldsymbol{h}$ is a nonnegative 1-by-N_t vector. This is, in fact, a rank-one nonnegative matrix factorization (NMF) [22] $\boldsymbol{B}=\boldsymbol{f}\boldsymbol{h}$ and we solve it using singular value decomposition (SVD). Finally, we use a Gaussian moving average filter to smooth $\boldsymbol{f}$ and $\boldsymbol{h}$ to obtain the estimation of piecewise-constant triggering densities. This is based on our assumption that g is smooth and can reduce the variance of our estimations. Our numerical experiments show that the regularization procedure described above leads to stable and robust estimations for synthetic and real-world data sets.

3.3 Triggering Matrix Estimation

In previous sections, we estimate the triggering density with the assumption that both $\boldsymbol{\mu}$ and $\boldsymbol{K}$ are given. In the univariate case, one can remove this assumption by adding two additional linear equations [40]. However, in multivariate case, because matrix $\boldsymbol{A}$ is depend on matrix $\boldsymbol{K}$, solving $\boldsymbol{\mu}$, $\boldsymbol{K}$ and g simultaneously is no longer a linear problem.

In order to solve this problem, we extend the cumulants (4), (5), (22) to the ST case for a fast estimation of $\boldsymbol{\mu}$ and $\boldsymbol{K}$. For a ST-Hawkes process with U sub-processes, we define its first, second and third cumulant as [11]

$$\Lambda^i dtdxdy = \mathbb{E}(dN^i_{t,x,y}), \tag{20}$$

$$C^{ij}dtdxdy = \int_{\tau,a,b\in\mathbb{R}^3} (\mathbb{E}(dN^i_{t,x,y}dN^j_{t+\tau,x+a,y+b} - \mathbb{E}(dN^i_{t,x,y})\mathbb{E}(dN^j_{t+\tau,x+a,y+b}), \tag{21}$$

$$\begin{aligned}\Gamma^{ijk}dtdxdy = \int_{\tau',a',b'\in\mathbb{R}^3}\int_{\tau,a,b\in\mathbb{R}^3} &(\mathbb{E}(dN^i_{t,x,y}dN^j_{t+\tau,x+a,y+b}dN^k_{t+\tau',x+a',y+b'})\\ &+2\mathbb{E}(dN^i_{t,x,y})\mathbb{E}(dN^j_{t+\tau,x+a,y+b})\mathbb{E}(dN^k_{t+\tau',x+a',y+b'})\\ &-\mathbb{E}(dN^i_{t,x,y}dN^j_{t+\tau,x+a,y+b})\mathbb{E}(dN^k_{t+\tau',x+a',y+b'})\\ &-\mathbb{E}(dN^i_{t,x,y}dN^k_{t+\tau,x+a,y+b})\mathbb{E}(dN^j_{t+\tau',x+a',y+b'})\\ &-\mathbb{E}(dN^j_{t+\tau,x+a,y+b}dN^k_{t+\tau',x+a',y+b'})\mathbb{E}(dN^i_{t,x,y})).\end{aligned} \tag{22}$$

Here $1 \leq i, j, k \leq U$ and τ, a and b are variables corresponding to t, x and y.

Cumulants can be numerically estimated from the ST data of events from each subprocess $Z^i = (t_k, x_k, y_k)_k, i = 1, ..., U$ on the ST bounded area $B \times [0, T]$. Here we simply assume that B is a rectangular with length X and width Y. We obtain the following the estimation formulas for (20), (21) and (22)

$$\hat{\Lambda}^i = \frac{1}{TXY} \sum_{\tau,a,b \in Z^i} = \frac{N^i_{T,X,Y}}{TXY}, \tag{23}$$

$$\hat{C}^{ij} = \frac{1}{TXY} \sum_{\tau,a,b \in Z^i} (N^j_{a+\tilde{X},b+\tilde{Y},\tau+H} - N^j_{a-\tilde{X},b-\tilde{Y},\tau-H} - 8\tilde{X}\tilde{Y}H\hat{\Lambda}^j), \tag{24}$$

$$\begin{aligned}\hat{\Gamma}^{ijk} =& \frac{1}{TXY} \sum_{\tau,a,b \in Z^i} (N^j_{a+\tilde{X},b+\tilde{Y},\tau+H} - N^j_{a-\tilde{X},b-\tilde{Y},\tau-H} - 8\tilde{X}\tilde{Y}H\hat{\Lambda}^j) \times \\ & (N^k_{a+\tilde{X},b+\tilde{Y},\tau+H} - N^k_{a-\tilde{X},b-\tilde{Y},\tau-H} - 8\tilde{X}\tilde{Y}H\hat{\Lambda}^k) \\ & - \frac{\hat{\Lambda}^i}{TXY} \sum_{\tau',a',b' \in Z^k} \sum_{\tau,a,b \in Z^j} (2H - |\tau - \tau'|)^+ (2\tilde{X} - |a - a'|)^+ (2\tilde{Y} - |b - b'|)^+ \\ & + 64(H\tilde{X}\tilde{Y})^2 \hat{\Lambda}^i \hat{\Lambda}^j \hat{\Lambda}^k \end{aligned} \tag{25}$$

via numerical approximations (see the Appendix B.3 in [1] for more details) of the cumulants on $[-\tilde{X}, \tilde{X}] \times [-\tilde{Y}, \tilde{Y}] \times [-H, H]$ assuming that the support of the triggering density is within this region. One also needs to symmetrize the approximated cumulants via $(\hat{C}^{ij} + \hat{C}^{ji})/2$ and $(2 * \hat{\Gamma^{ijj}} + \hat{\Gamma^{jii}})/3$ because $\Gamma^{iji} = \Gamma^{iij}$ and $C^{ij} = C^{ji}$. Finally we can plug the approximated cumulants into (7) to estimate $\boldsymbol{\mu}$ and $\boldsymbol{K}$. The error function (7) is a non-convex polynomial and similar to the loss function of a multilayer neural network. As a result, stochastic gradient descend (SGD) with acceleration (e.g. Adam [20] or AdaGrad [13]) can be used to minize the error function. The normalization term κ is $\kappa = \frac{\|\hat{\boldsymbol{\Gamma}}^{\mathbf{c}}\|_2^2}{\|\hat{\boldsymbol{C}}\|_2^2 + \|\hat{\boldsymbol{\Gamma}}^{\mathbf{c}}\|_2^2}$ based on the thery of GMM [1]. The ratio between the support of the triggering density and the ST bounded area $B \times [0, T]$ matters for the consistency of the GMM [1]. Usually for specific applications such as social network reconstruction, $B \times [0, T]$ is much larger than the support of the triggering density, which guarantees the consistency of the GMM estimation.

3.4 Consistency Guarantee

The consistency of maximum likelihood estimates [31] or GMM estimates [1] is guaranteed by general theoretical results. Here we note that our proposed method, as a combination of GMM and MLE, also yields consistent estimates.

First, as background, note that in [31], Ogata shows the MLE of the full vector of parameters is, under quite general conditions, consistent. Also, if only some of the parameters are to be estimated and others, such as in this instance $\boldsymbol{K}$ and $\boldsymbol{\mu}$, are known exactly, then again one may consider the parameter vector to be only those parameters being estimated, and again [31] implies the estimated ones will be consistent. However, we are considering the case where $\boldsymbol{K}$ and $\boldsymbol{\mu}$ are not known but are estimated consistently via GMM, and then the other parameters are estimated by MLE. To the best of our knowledge, this case has not been studied previously, and the result does not immediately follow from the theorems in [31]. We show that $\hat{\boldsymbol{\beta}}$ will be consistent in this case. We will assume the same assumptions as in [31, 1]. While the proof does not follow directly from the theorem, it can be proven in essentially the same manner.

Let Θ denote the full vector of parameters, including $\boldsymbol{K}$ and $\boldsymbol{\mu}$. Let Θ_0 denote the true value of Θ. Let $\boldsymbol{U}$ denote a neighborhood of Θ_0. Let $\boldsymbol{K'}$ and $\boldsymbol{\mu'}$ denote the GMM estimates of $\boldsymbol{K}$ and $\boldsymbol{\mu}$. Let $\Theta = (\boldsymbol{K}, \boldsymbol{\mu}, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the vector of other parameters estimated by MLE. Let $\hat{\boldsymbol{K}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\beta}}$ be the MLEs of these parameters.

Theorem 1. $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ in probability as $T \to \infty$.

Proof. Let L denote the loglikelihood divided by T. Thus L depends on T but we will suppress this here. Let Θ_1 denote the supremum over $\boldsymbol{U}^c$ of L, i.e. the MLE outside of $\boldsymbol{U}$.

We are given that $(\boldsymbol{K'}, \mu') \to (\boldsymbol{K}, \mu)$ in probability as $T \to \infty$. Thus, ultimately $(\boldsymbol{K'}, \mu')$ are in $\boldsymbol{U}$ with probability going to 1.

$L(\Theta_1) \to \mathbb{E}(L(\Theta_1))$ and $L(\Theta_0) \to \mathbb{E}(L(\Theta_0))$, where this convergence is in probability as $T \to \infty$ and is uniform in Θ. This follows from exactly the same logic as in the proof of Theorem 2 of [31].

Similarly, following exactly as in the proof on p253 of [31], $\mathbb{E}(L(\Theta_0)) > \mathbb{E}(L(\Theta_1))$ and

further, for sufficiently large T, there exists $\epsilon > 0$ so that $|\mathbb{E}(L(\Theta)) - \sup_{\Theta \notin U} \mathbb{E}(L(\Theta))| > \epsilon/2$. This follows from the assumptions in [31], particularly the assumption that λ is uniformly bounded away from 0.

Therefore, since $(\boldsymbol{K}', \mu')$ are ultimately in $\boldsymbol{U}$, for sufficiently large T, $\mathbb{E}(L(\Theta_0))$ and therefore $L(\Theta_0)$ are also maximized within U with probability going to 1. More specifically, for any $\epsilon > 0$, there is $\delta > 0$ and sufficiently large T so that

$$\begin{aligned}
P(\hat{\boldsymbol{\beta}} \notin \boldsymbol{U}) &= P\{\sup_{\boldsymbol{U}^c} L(\Theta) \geq \sup_{\boldsymbol{U}} L(\Theta)\} \\
&\leq P\{L(\Theta_1) \geq L(\Theta_0)\} \\
&\leq P\{L(\Theta_1) - \mathbb{E}(L(\Theta_1)) \geq \delta\} + P\{\mathbb{E}(L(\Theta_1)) - \mathbb{E}(L(\Theta_0)) > -2\delta\} \\
&\quad + P\{\mathbb{E}(L(\Theta_0)) - L(\Theta_0) \geq \delta\} \\
&\leq \epsilon/2 + 0 + \epsilon/2 \\
&= \epsilon.
\end{aligned}$$

3.5 Computational Complexity

The state-of-the-art cumulants-based method (NPHC) [1] for temporal triggering density estimation has a complexity of $O(NU^2 + N_{iter}U^3)$, where N_{iter} is the number of iterations. Our method has a similar complexity $O(NU^2 + N_{iter}U^3 + (N_r N_t)^3)$ as NPHC since the calculation time of spatialtemporal cumulants is just a constant multiple of temporal cumulants. The additional calculation for triggering density estimation is usually neglectable because N_r, N_t are small constants and $\boldsymbol{A}$ is usually sparse. For EM-type algorithm (EM) [23], the complexity is $O(N_{iter}N^3U^2)$ [1]. With some clever implementation or in some special cases (e.g. temporal Hawkes process with an exponential triggering density), one can reduce this to $O(N^2)$ or better.

Our method outperforms EM when $N \gg U$. Moreover, in many cases, we find that our method is even faster than NPHC. This seems impossible since our method needs to process spatial data in addition to the timestamp. However, for ST data, there are many event pairs that are close in time (within the support of the temporal triggering density) while spatially separated from each other (outside the support of the spatial triggering density). Temporal-only model such as NPHC will calculate these events pairs during the

estimation of cumulants. This might cause false positives in causal inference. Our method, on the other hand, uses spatial information to exclude these events. It seems that, for a majority of data sets we examined, this effect is very significant and our method can be several time faster than NPHC.

4 Numerical Examples

In this section, we compare our method (we call it STHC—ST Hawkes cumulants—throughout this section) with other popular estimation methods for multivariate Hawkes processes on various data sets. For a thorough comparison, we consider both simulation data and real-world social network data. First, we simulate multiple synthetic data sets with different sizes, triggering matrices and triggering densities. These data sets with ground-truth information allow us to examine different methods in detail. Then for real-world applications, we further evaluate the performance of these methods on the task of network reconstruction for multiple location-based social network check-in data sets. Moreover, our method directly estimates spatial and temporal triggering densities, which provides a useful tool for the study of ST dynamics among these check-in events. All the experiments are conducted on a single machine with a NVIDIA 970 GPU (4 GB memory), 4-core Intel i7-6700K CPU (4.20 GHz), and 16 GB of RAM.

4.1 Synthetic Data

Our synthetic data sets are generated using Algorithm 3 in [42], which is based on the clustering representation of Hawkes process. We simulate various ST-Hawkes processes and use them to evaluate our method (STHC), the state-of-the-art temporal cumulants method (NPHC) and EM-type Algorithm (EM). The details about the simulation and preprocessing are described in Appendix A. Here we define some error measurements used in this section.

- *Relative error* between the estimated triggering matrix $\hat{\boldsymbol{K}}$ and the ground-truth ma-

trix $\boldsymbol{K}$:

$$\text{RelErr}(\boldsymbol{K}, \hat{\boldsymbol{K}}) = \frac{1}{U^2} \sum_{u,v} \left(\frac{|K_{uv} - \hat{K}_{uv}|}{|K_{uv}|} \mathbb{1}_{K_{uv} \neq 0} + |\hat{K}_{uv}| \mathbb{1}_{K_{uv}=0} \right)$$

- *Mean squared error* (MSE) between the estimated triggering densities (temporal $\hat{h}(t)$, spatial $\hat{f}(r)$ and combined $\hat{g}(r,t)$) and the ground-truth triggering densities (temporal $h(t)$, spatial $f(r)$ and combined $g(r,t)$):

$$\text{MSE}_r = \frac{1}{N_r} \sum_{i=1}^{N_r} (f_i - \hat{f}_i)^2, \ \text{MSE}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} (h_i - \hat{h}_i)^2, \ \text{MSE}_\beta = \frac{1}{N_r N_t} \sum_{i=1}^{N_r} \sum_{j=1}^{N_t} (g_{ij} - \hat{g}_{ij})^2$$

Here $\hat{g}_{ij} = \boldsymbol{B}(i,j)$ is the discrete estimation of the triggering density on a 2-D grid of size $N_t \times N_r$ and g_{ij} are the ground-truth values of the triggering density on the grid. $\hat{h}_i = \hat{\boldsymbol{h}}(i)$ and $\hat{f}_i = \hat{\boldsymbol{f}}(i)$ are from the NMF decomposition of $\boldsymbol{B}$, and $h_i = \boldsymbol{h}(i)$ and $f_i = \boldsymbol{f}(i)$ are the ground-truth values of the temporal and spatial triggering densities on the grid accordingly.

Triggering Density estimation

We first compare our methods with EM in terms of the triggering density estimation accuracy (NPHC does not estimate triggering densities). The simulation data with $2,587$ events is from a ST Hawkes process with $U = 1$, exponential triggering density in time and Gaussian in space. We get a good estimation of the triggering density $f(r)$ ($\text{MSE}_r = 0.001662$), $h(t)$ ($\text{MSE}_t = 0.02876$) in Figure 1 and the overall estimation for $\boldsymbol{\beta} = (g_{mn})_{k(m,n)}$ ($\text{MSE}_\beta = 0.03400$). This is a relatively small data sets so that we can use EM for ST Hawkes (ST-EM, see [42]) estimation. For ST-EM, we get $f(r)$ ($\text{MSE}_r = 0.01485$), $h(t)$ ($\text{MSE}_t = 0.004058$) and $\boldsymbol{\beta}$ ($\text{MSE}_\beta = 0.2533$). Our method is faster (see Table 1) and overall more accurate.

Triggering matrix

Then we evaluate the ability of our model to recover the triggering matrix $\boldsymbol{K}$. This is important for many applications such as network reconstruction and causal inference. On our existing architecture, the ST-EM method runs out of memory. Instead, we use EM and NPHC implementations in the tick package [2] for the following comparisons.

Figure 1: The estimation results of STHC on $U = 1$ data. Ground truth spatial triggering density $f(r)$ as red triangles and estimated triggering density as blue circles (left). Temporal triggering density $h(t)$ as red triangles and estimated triggering density as blue circles (right).

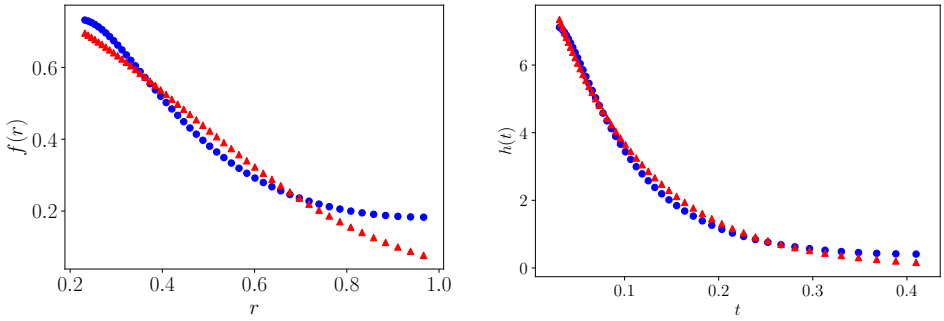
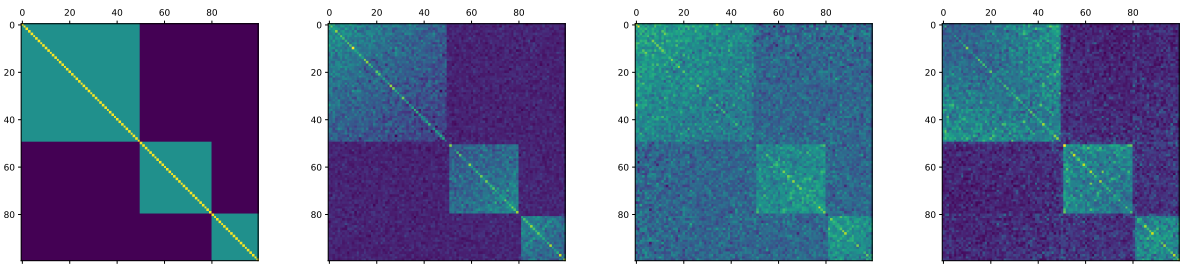


Figure 2: Ground truth $\boldsymbol{K}$ matrix, STHC , NPHC result and EM estimation results (from left to right).



We simulate a ST-Hawkes process with $U = 100$ and a symmetric $\boldsymbol{K}$ matrix (see Figure 2) because our network reconstruction data sets mainly have undirected social networks. We achieve a relative error of 0.1080. In the same setting, we get a relative error of 0.1626 for NPHC and 0.1459 for EM. The improvement in computation time (see Table 1) is significant.

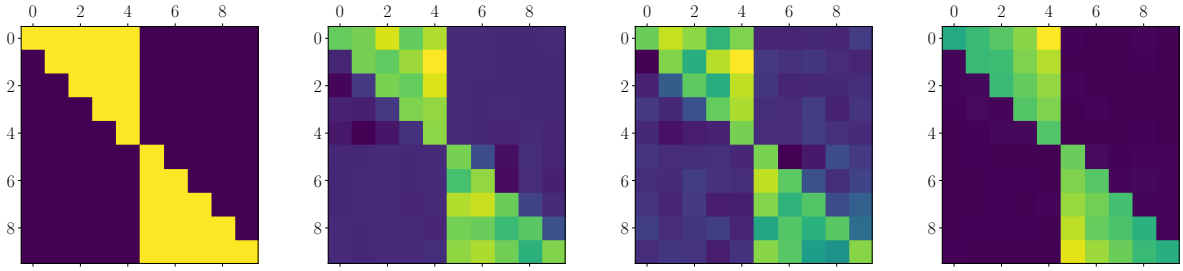
Combined estimation

Now we combine the two steps together and give a complete estimation of ST-Hawkes processes. We simulate a ST-Hawkes process with $U = 10$ and $179,176$ events in total. From the results in Figure 3 and Table 1, STHC gives very fast and also accurate estimations

Table 1: The computation time for different methods on synthetic data sets. Here the time is in second.

	STHC	NPHC	EM
$U = 1$	0.165528	-	4.643132
$U = 10$	1.073085	1.093068	4.707377
$U = 100$	2.608996	4.174796	43.781988

Figure 3: Ground truth $\boldsymbol{K}$ matrix, STHC, NPHC and EM estimation results (from left to right).



(RelErr=0.02901) comparing to NPHC (RelErr=0.04899) and EM (RelErr=0.03269). We then threshold $\hat{\boldsymbol{K}}$ with $\epsilon = 0.01$ to remove noise. Using $\hat{\boldsymbol{K}}, \hat{\boldsymbol{\mu}}$, we get a good estimation of the triggering density $f(r)$ and $h(t)$ in Figure 4 with $\text{MSE}_r = 0.002381$, $\text{MSE}_t = 0.06664$ and $\text{MSE}_\beta = 0.1067$ while EM has a much worse MSE ($\text{MSE}_t = 0.9512$) since it does not consider spatial information.

Combined estimation with different triggering densities

We modify the above $U = 10$ data set via replacing the ST triggering density with different functions. We first get accurate estimations of $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{\mu}}$. Given $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{\mu}}$, we then estimate the triggering density in space and time (See Figure 5). The results are summerized in Table 2. Specifically, we consider Pareto triggering density in time, uniform triggering density in time, power-law triggering density in space and uniform triggering density in space. See Appendix A for more details on generating these synthetic data sets.

Figure 4: The estimation results of STHC on $U = 10$ data. Ground truth spatial triggering density $f(r)$ as red triangles and estimated triggering density as blue circles (left). Temporal triggering density $h(t)$ as red triangles and estimated triggering density as blue circles (right).

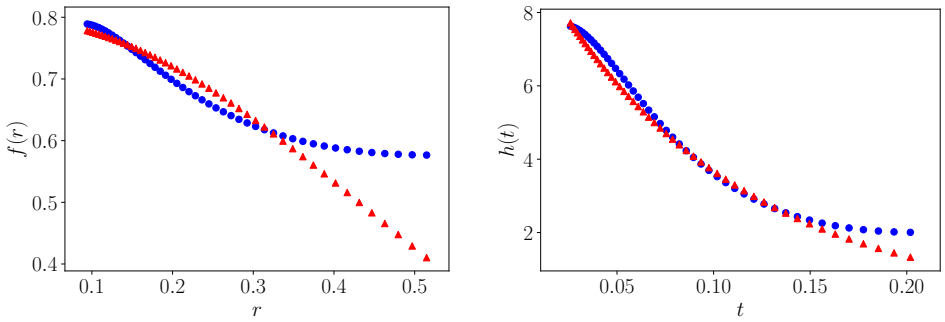


Figure 5: The estimation results of STHC on $U = 10$ data with a Pareto triggering density in time, a uniform triggering density in time, a power-law triggering density in space and a uniform triggering density in space (from left to right). (Top) Ground truth spatial triggering density $f(r)$ as red triangles and estimated triggering density as blue circles. (Bottom) Temporal triggering density $h(t)$ as red triangles and estimated triggering density as blue circles.

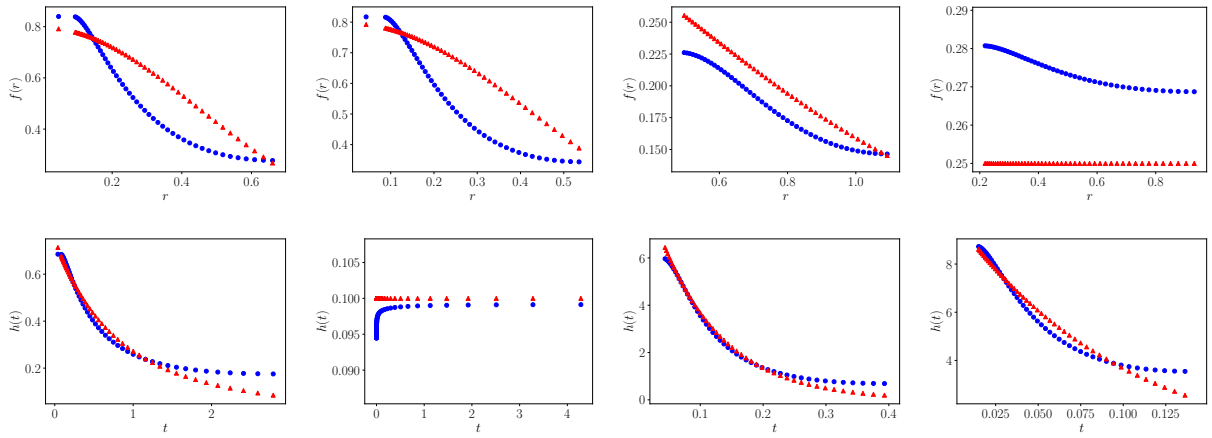


Table 2: Error measures for STHC on $U = 10$ data sets with different triggering densities.

	MSE_r	MSE_t	$\text{RelErr}(\boldsymbol{K}, \hat{\boldsymbol{K}})$
Pareto in time	0.01244	0.0009966	0.02784
Uniform in time	0.01320	1.296×10^{-5}	0.09306
Power-law in space	0.0003904	0.04463	0.0409
Uniform in space	0.0006231	0.1294	0.04552

4.2 Location-based Social Network Reconstruction

In many situations, network data are incomplete and it may not be possible to directly observe the hidden relationships between nodes. Our task of network reconstruction is to uncover the ground-truth friendship network among social media users using only the information of each user's check-ins.

The Gowalla and Brightkite data sets, collected in [9], are both from location-based social-media websites in which users share their locations by checking in. Gowalla has a "friendship" network with 196,591 users, 950,327 edges, and a total of 6,442,890 check-ins of these users between February 2009 and October 2010. Brightkite's "friendship" network consists of 58,228 nodes and 214,078 edges, and a total of 4,491,143 check-ins over the period of Apr. 2008 - Oct. 2010. Each check-in record includes the latitude and longitude coordinates, a user ID and the time (with a precision of one second). Similar to the Facebook "friendship" network, both the Gowalla and Brightkite friendship networks are undirected and unweighted. We study several subnetworks (Gowalla-SF, Brightkite-LA, Gowalla-CHI, and Brightkite-SD) within these data sets; see Appendix B for details.

We model the ST check-ins of each user within a subnetwork as events of one subprocess within a multivariate ST-Hawkes process. Then we infer relationships between these users (i.e. infer adjacency matrix) from the triggering matrix $\boldsymbol{K}$, which uncover the macro-scale causality between users (subprocesses). Our assumption here is that this causality information reflects actual friendship connections. We compare our method (STHC) with

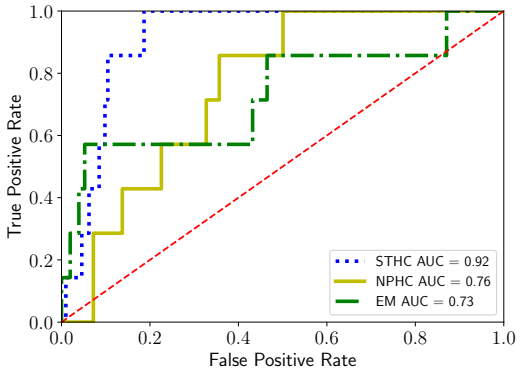
NPHC and EM in terms of how well the reconstructed networks match the ground-truth friendships. With the prior information that friendship networks are undirected, we first symmetrize the inferred triggering matrix (via $\tilde{\boldsymbol{K}} = \left(\hat{\boldsymbol{K}} + \hat{\boldsymbol{K}}^T\right)/2$) to obtain the estimated weighted adjacency matrix. Then the network reconstruction becomes a binary classification problem with the probability $\propto \tilde{\boldsymbol{K}}$. Given the ground-truth binary adjacency matrix, we calculate the corresponding receiver operating characteristic (ROC) curves and the area under the curve (AUC) to evaluate the results.

The performances of different methods are examined on various subnetworks with different sizes. Our STHC method consistently outperforms other methods with more than 20% improvement in terms of the AUC in Figure 6. The improvement is mainly from the ability of our method to exclude false-positive connections. We show an example of network reconstruction results of Brightkite-SD in Figure 7. For the computation time (See Table 3), STHC scales better than NPHC in all data sets, as explained in Section 3.5. EM has the worst scaling due to its super-linear complexity. Finally, we estimate spatial and temporal triggering densities for these subnetworks and plot them in Figure 8. The spatial triggering densities for different subnetworks have similar shapes with a cut-off around 10^{-4}. This could come from the fact that the check-in location is usually fixed for a point of interest (POI, such as shop/cafe/gym). The triggering density also implies that the spatial triggering effects between users have a short radius, which mainly occur when they visit the same POI. These temporal triggering densities also share the same trend. The triggering effects only peak a few hours after the event time. This is also observed in other data sets, such as the insurgency activity in Iraq [23].

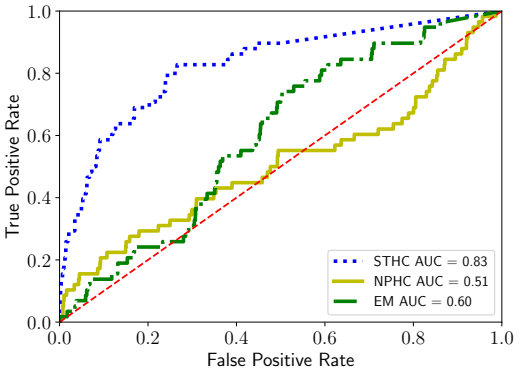
5 Conclusion

We present a novel inference approach of ST-Hawkes processes and it is the most efficient and accurate method comparing with other popular estimation methods, according to the numerical experiments presented above. Moreover, this approach is successfully applied to network reconstruction problems and leads to promising applications for the inference of causality and social interactions.

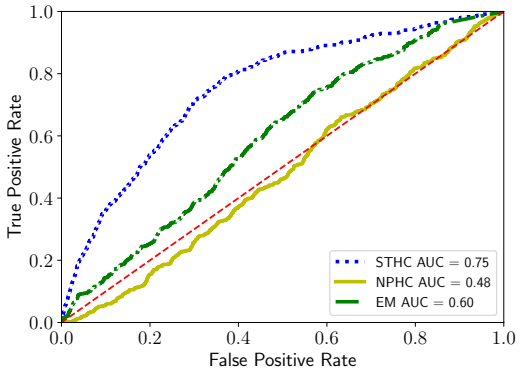
Figure 6: ROC curves of different methods (STHC, NPHC and EM) on subnetworks in Gowalla and Brightkite data sets. The dashed line (red) is from random guess.



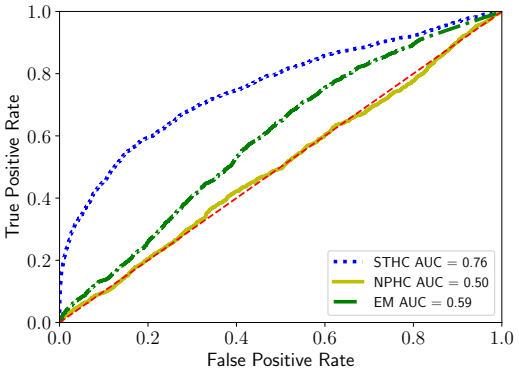
(a) Brightkite-SD



(b) Gowalla-CHI

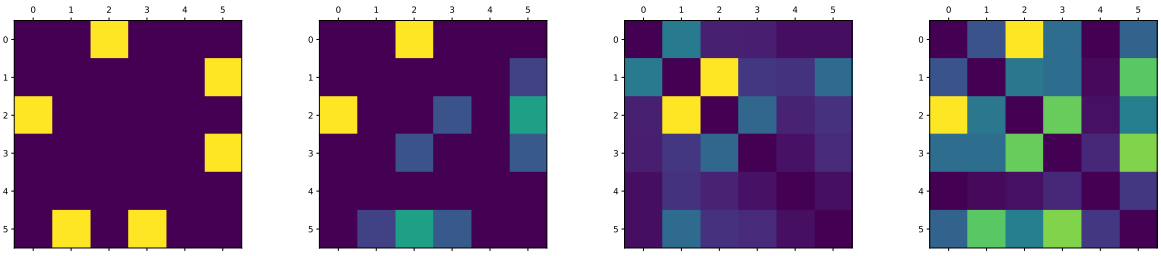


(c) Brightkite-LA



(d) Gowalla-SF

Figure 7: Friendship network reconstruction using different methods on Brightkite-SD. Here we zoom in to show a subgraph within the Brightkite-SD network.

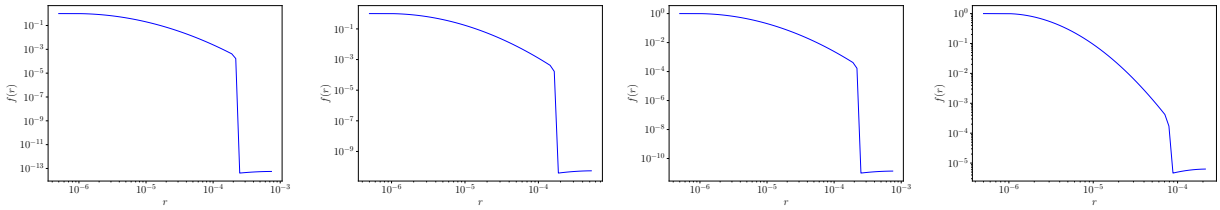


(a) Groundtruth (b) STHC (c) NPHC (d) EM

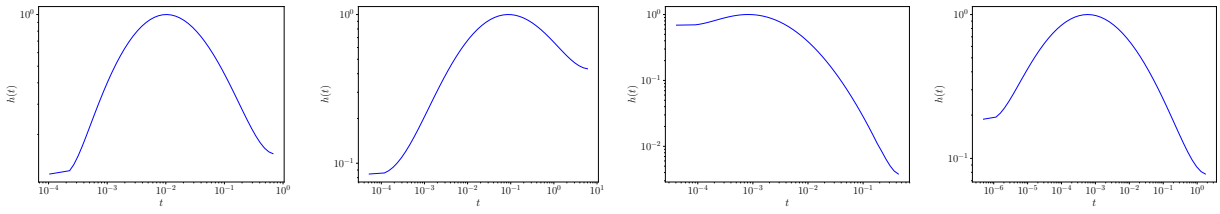
Table 3: The computation time for different methods on Gowalla and Brightkite data sets. Here the time is in second.

	STHC	NPHC	EM
Brightkite-SD	0.271304	2.035561	2.252009
Gowalla-CHI	2.978064	3.869652	15.474624
Brightkite-LA	3.976395	7.001311	36.357789
Gowalla-SF	40.754037	76.514422	180.918273

Figure 8: Estimated spatial and temporal triggering densities of our method on Gowalla and Brightkite data sets. The plot is in log-log scale and we normalize the triggering density for easy comparison.



(a) Spatial triggering densities for Brightkite-SD, Gowalla-CHI, Brightkite-LA and Gowalla-SF.



(b) Temporal triggering densities for Brightkite-SD, Gowalla-CHI, Brightkite-LA and Gowalla-SF.

A point that should be stressed is that we make a few model assumptions to simplify the estimation procedure. To recapitulate, we assume a constant background rate in space and no boundary effect for events outside the area we studied. For more general spatial background (inhomogeneous) distribution, one can approximate it using a piece-wise constant function in space by dividing events into spatial grids. Essentially for each grid, we still have a uniform background for estimation and then combine them together. For applications on large areas with an inhomogeneous background, we expect a piece-wise constant or covariate-based background rate to achieve even better results [40]; and incorporating boundary effects helps to remove bias in the estimation of the background rate and triggering densities [35]. Moreover, the current regularization method can be extended to a more general case to utilize the smoothness proprieties of triggering densities.

Finally, while we are focusing on the general case of multivariate ST-Hawkes processes, the current method can be very useful for the estimation of univariate models. The regularization improves the stability and robustness of the analytic method in [40]. This makes it possible to apply univariate models to the study of large data sets in seismology, epidemiology, and criminology.

Acknowledgement

This work was supported by the City of Los Angeles Gang Reduction Youth Development Project and by NSF grant DMS-1737770. BY gratefully acknowledges the fellowship support of the National Institute of Justice (NIJ) under award number 2018-R2-CX-0013.

References

[1] M. Achab, E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy. Uncovering causality from multivariate hawkes integrated cumulants. *The Journal of Machine Learning Research*, 18(1):6998–7025, 2017.

[2] E. Bacry, M. Bompaire, S. Gaïffas, and S. Poulsen. tick: a Python library for statistical

learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, July 2017.

[3] E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

[4] E. Bacry and J.-F. Muzy. First-and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

[5] E. Balderama, F. P. Schoenberg, E. Murray, and P. W. Rundel. Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, 107(498):467–476, 2012.

[6] P. J. Brantingham, B. Yuan, N. Sundback, F. P. Schoenberg, A. L. Bertozzi, J. Gordon, J. Leap, K. Chan, M. Kraus, S. Malinowski, et al. Does violence interruption work? 2018.

[7] D. R. Brillinger, P. M. Guttorp, F. P. Schoenberg, A. H. El-Shaarawi, and W. W. Piegorsch. Point processes, temporal. *Encyclopedia of Environmetrics*, 3:1577–1581, 2002.

[8] S. Chen, A. Shojaie, E. Shea-Brown, and D. Witten. The multivariate hawkes process in high dimensions: beyond mutual excitation. *arXiv preprint arXiv:1707.04928*, 2017.

[9] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[10] D. J. Daley and D. Vere-Jones. An introduction to the theory of point processes. vol. i. probability and its applications, 2003.

[11] 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.

[12] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 219–228. ACM, 2015.

[13] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[14] M. Eichler, R. Dahlhaus, and J. Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

[15] M. Farajtabar, Y. Wang, M. G. Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, pages 1954–1962, 2015.

[16] 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. *Journal of the American Statistical Association*, 111(514):564–584, 2016.

[17] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[18] E. C. Hall and R. M. Willett. Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 62(7):4327–4346, 2016.

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

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[21] E. L. Lai, D. Moyer, B. Yuan, E. Fox, B. Hunter, A. L. Bertozzi, and P. J. Brantingham. Topic time series analysis of microblogs. *IMA Journal of Applied Mathematics*, 81(3):409–431, 2016.

[22] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[23] E. Lewis and G. Mohler. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.

[24] S. Linderman and R. Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

[25] B. Mark, G. Raskutti, and R. Willett. Network estimation from point process data. *IEEE Transactions on Information Theory*, 2018.

[26] D. Marsan and O. Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

[27] G. O. Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497, 2014.

[28] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[29] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

[30] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM review*, 40(3):636–666, 1998.

[31] Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.

[32] Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

[33] W. L. Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.

[34] M. D. Porter, G. White, et al. Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124, 2012.

[35] A. Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

[36] F. P. Schoenberg. On non-simple marked point processes. *Annals of the Institute of Statistical Mathematics*, 58(2):223–233, 2006.

[37] F. P. Schoenberg. Facilitated estimation of etas. *Bulletin of the Seismological Society of America*, 103(1):601–605, 2013.

[38] F. P. Schoenberg. Comment on "a review of self-exciting spatio-temporal point processes and their applications" by alex reinhart. *Statist. Sci.*, 33(3):325–326, 08 2018.

[39] F. P. Schoenberg, D. R. Brillinger, and P. Guttorp. Point processes, spatial-temporal. *Wiley StatsRef: Statistics Reference Online*, 2014.

[40] F. P. Schoenberg, J. S. Gordon, and R. J. Harrigan. Analytic computation of nonparametric marsan–lengliné estimates for hawkes point processes. *Journal of Nonparametric Statistics*, pages 1–16, 2018.

[41] A. Veen and F. P. Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

[42] B. Yuan, H. Li, A. L. Bertozzi, P. J. Brantingham, and M. A. Porter. Multivariate spatiotemporal hawkes processes and network reconstruction. *arXiv preprint arXiv:1811.06321*, 2018.

[43] J. Zhuang, Y. Ogata, and D. Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.

Appendices

A Simulation Data

$U = 1$ Data

We simulate a univariate ST Hawkes process with $K = 1/6$, $\mu = 0.01$, $T = 2.1 \times 10^5$, $X, Y \in (0, 10)$, $f(r) = \frac{1}{2\pi\sigma^2} \exp(-r^2/2\sigma^2)$ ($\sigma^2 = 0.2$) and $h(t) = \omega \exp(-\omega t)$ ($\omega = 10$). The regularization parameter $\alpha = 0.5$.

$U = 100$ Data

Using the same triggering densities, this data set has the following parameters: $U = 100$, the background rate $\boldsymbol{\mu} = (0.01, ..., 0.01)$. $T = 10^5$, $X, Y \in (0, 10)$, $\sigma^2 = 0.2$ and $\omega = 10$ with $172,943$ events. For the triggering matrix in Figure 2, each yellow pixel is $1/20$, cyan pixel is $1/40$ and dark pixel is 0.

$U = 10$ Data

With the same densities, the parameters are $U = 10$, $\boldsymbol{\mu} = (0.01, ..., 0.01)$, $T = 1e6$, $X, Y \in (0, 10)$, $\sigma^2 = 0.2$, $\omega = 10$ and $\boldsymbol{K}$ is shown in Figure 3. Here each yellow pixel is $1/6$ and dark pixel is 0. The regularization parameter $\alpha = 0.55$.

$U = 10$ Data with a Pareto Triggering Density in Time

We keep the same parameters as the $U = 10$ above. The changes on the densities are on the temporal density $h(t) = (p-1)c^{p-1}/(t+c)^p$ with $c = 2$ and $p = 2.5$ and the same spatial triggering density with $\sigma^2 = 0.1$. The regularization parameter $\alpha = 0.38$.

$U = 10$ Data with a Uniform Triggering Density in Time

Similar to the section above, here we change the temporal densities to be uniform $h(t) = 0.1$ and the spatial triggering density with $\sigma^2 = 0.1$. The regularization parameter $\alpha = 0.4$. We threshold the estimated $\tilde{\boldsymbol{K}}$ with $\epsilon = 0.01$ to remove noise.

$U = 10$ Data with a Power-law Triggering Density in Space

Similarly, we use the power-law denisity $f(r) = \frac{1}{(r^2+1)^2}$ in space and the exponential triggering density in time with $\omega = 10$. The regularization parameter $\alpha = 0.28$. We threshold the estimated $\tilde{\boldsymbol{K}}$ with $\epsilon = 0.02$ to remove noise.

$U = 10$ Data with a Uniform Triggering Density in Space

Given the same parameters as above, we change the spatial density to $f(r) = 0.25$ and keep the exponential triggering density in time with $\omega = 10$. The regularization parameter $\alpha = 0.36$. We threshold the estimated $\tilde{\boldsymbol{K}}$ with $\epsilon = 0.01$ to remove noise.

B Gowalla and Brightkite data sets

In this section, we describe the preprocessing procedure for Gowalla and Brightkite data sets. We focus on various local friendship subnetworks within different U.S. cities, including San Diego (SD), Chicago (CHI), Los Angeles (LA) and San Fransico (SF). They have diverse network sizes and ST patterns within the same time period.

Brightkite-SD

We study check-ins in SD for Brightkite data set. We use a bounding box (with a north latitude of 33.1142, a south latitude of 32.5348, an east longitude of -116.9058, and a west longitude of -117.2824)[1] to locate check-ins in SD. We consider "active" users, who have more than 300 check-ins during the period. This gives us a small subnetwork with 25 "active" users and a total of $13,760$ check-ins in SD.

Gowalla-CHI

We apply the same procedure as in B on the Gowalla check-in data for CHI. The bounding box for CHI has a north latitude of 42.0229, a south latitude of 41.6446, an east longitude of -87.5245, and a west longitude of -87.9395. After selecting only active users (with more

[1] We obtain latitude and longitude coordinates from https://www.flickr.com/places/info.

than 100 check-ins) users, we have a medium-sized subnetwork with 96 users and $27,326$ check-ins.

Brightkite-LA

We apply the same procedure as in B on the Brightkite check-in data in LA. The bounding box for LA has a north latitude of 34.34, a south latitude of 33.70, an east longitude of -118.16, and a west longitude of -118.67. After selecting only active users (with more than 150 check-ins) users, we have a medium-sized subnetwork with 168 users and $89,127$ check-ins.

Gowalla-SF

We apply the same procedure as in B on the Gowalla check-in data in SF. The bounding box for SF has a north latitude of 37.93, a south latitude of 37.64, an east longitude of -122.28, and a west longitude of -123.17. After selecting only active users (with more than 65 check-ins) users, we have a large subnetwork with 515 users and $102,673$ check-ins.