Learning Network of Multivariate Hawkes Processes: A Time Series Approach

Jalal Etesami
Dep. of Industrial &
Enterprise Systems Eng.
Coordinated Science Lab.
UIUC, Urbana, IL 61801
etesami2@illinois.edu

Negar Kiyavash
Dep. of Industrial &
Enterprise Systems Eng.
Dep. of Electrical &
Computer Eng.
Coordinated Science Lab.
UIUC, Urbana, IL 61801
kiyavash@illinois.edu

Kun Zhang
Dep. of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213
kunz1@andrew.cmu.edu

Kushagra Singhal
Dep. of Electrical &
Computer Eng.
Coordinated Science Lab.
UIUC, Urbana, IL 61801
ksingha2@illinois.edu

Abstract

Learning the influence structure of multiple time series data is of great interest to many disciplines. This paper studies the problem of recovering the causal structure in network of multivariate linear Hawkes processes. In such processes, the occurrence of an event in one process affects the probability of occurrence of new events in some other processes. Thus, a natural notion of causality exists between such processes captured by the support of the excitation matrix. We show that the resulting causal influence network is equivalent to the Directed Information graph (DIG) of the processes, which encodes the causal factorization of the joint distribution of the processes. Furthermore, we present an algorithm for learning the support of excitation matrix of a class of multivariate Hawkes processes with exponential exciting functions (or equivalently the DIG). The performance of the algorithm is evaluated on synthesized multivariate Hawkes networks as well as a stock market and MemeTracker real-world dataset.

1 INTRODUCTION

In many disciplines, including biology, economics, social sciences, and computer science, it is important to learn the structure of interacting networks of stochastic processes. In particular, succinct representation of the causal interactions in the network is of interest.

A lot of studies in the causality fields focus on causal discovery from time series. To find causal relations from time series, one may fit vector autoregressive models on the time series, or more generally, evaluate the causal influences with transfer entropy [22] or directed information [19]. This paper considers learning causal structure for a specific type of time series, multivariate linear Hawkes process [8]. Hawkes processes were originally motivated by the quest for good statistical models for earthquake occurrences. Since then, they have been successfully applied to seismology [15], biology [21], criminology [13], computational finance [5, 12, 14], etc. It is desirable to develop specific causal discovery methods for such processes and study the properties of existing methods in this particular scenario.

In multivariate or mutually exciting point processes, occurrence of an event (arrival) in one process affects the conditional probability of new occurrences, i.e., the *intensity* function of other processes in the network. Such interdependencies between the intensity functions of a linear Hawkes process are modeled as follows: the intensity function of processes j is assumed to be a linear combination of different terms, such that each term captures only the effects of one other process (See Section 2.1). Therefore, a natural notion of functional dependence (causality) exists among the processes in the sense that in linear mutually exciting processes, if the coefficient pertaining to the effects of process i is non-zero in the intensity function of process j, we know that process i is influencing process j. This dependency is captured by the support of the excitation matrix of the network. As a result, estimation of the excitation (kernel) matrix of multivariate processes is crucial both for learning the structure of their causal network and for other inference tasks and has been the focus of research. For instance, maximum likelihood estimators were proposed for estimating the parameters of excitation matrices with exponential and Laguerre decay in [16, 25]. These estimators depend on existence of i.i.d. samples. However, often we do not have access to i.i.d. samples when analyzing time series. Second-order statistics of the multivariate Hawkes processes were used to estimate the kernel matrix of a subclass of multivariate Hawkes processes called symmetric Hawkes processes [1]. Utilizing the branching property of the Hawkes processes, an expectation maximization algorithm was proposed to estimate the excitation matrix in [10].

We aim to investigate efficient approaches to estimation of excitation matrix of Hawkes processes from time series that does not require i.i.d. samples and investigate how the concept of causality in such processes is related to other established approaches to analyze causal effects in time series.

1.1 SUMMARY OF RESULTS AND ORGANIZATION

Our contribution in this paper is two fold. First, we prove that for linear multivariate Hawkes processes, the causal relationships implied by the excitation matrix is equivalent to a specific factorization of the joint distribution of the system called *minimal generative model*. Minimal generative models encode causal dependencies based on a generalized notion of Granger causality, measured by causally conditioned directed information [20]. One significance of this result is that it provides a surrogate to directed information measure for capturing causal influences for Hawkes processes. Thus, instead of estimating the directed information, which often requires estimating a high dimensional joint distribution, it suffices to learn the support of the excitation matrix. Our second contribution is indeed providing an estimation method for learning the support of excitation matrices with exponential form using second-order statistics of the Hawkes processes.

Our proposed learning approach, in contrast with the previous work [1, 24], is not limited to symmetric Hawkes processes. In a symmetric Hawkes process, it is assumed that the Laplace transform of the excitation matrix can be factored into product of a diagonal matrix and a constant unitary matrix. Moreover, it is assumed that the expected values of all intensities are the same. A numerical method to approximate the excitation matrix from a set of coupled integral equations was recently proposed in [3]. Our approach is based on an exact analytical solution to find the excitation matrix. Interestingly, the exact approach turns out to be both more robust and less expensive in terms of complexity compared to the numerical method of [3].

The rest of this paper is organized as follows. Background material, some definitions, and the notation are presented in Section 2. Specifically, therein, we formally introduce multivariate Hawkes processes and directed information graphs. In Section 3, we establish the connection between the excitation matrix and the corresponding DIG. In Section 4, we propose an algorithm for learning the excitation matrix or equivalently the DIG of a class of stationary multivariate linear Hawkes processes. Section 5 illustrates the performance of the proposed algorithm in inferring the causal structure in a network of synthesized mutually exciting linear Hawkes processes and in stock market. Finally, we conclude our work in Section 6.

2 PRELIMINARY DEFINITIONS

In this Section we review some basic definitions and our notation. We denote random processes by capital letters and a collection of m random processes by $\underline{X}_{[m]} = \{X_1, ..., X_m\}$, where $[m] := \{1, ..., m\}$. We denote the ith random process at time t by $X_i(t)$, the random process X_i from time s up to time t by $X_{i,s}^t$, and a subset $\mathcal{K} \subseteq [m]$ of random process up to time t by $\underline{X}_{\mathcal{K}}^t$. The Laplace transform and Fourier Transform of X_i are denoted, respectively by

$$\begin{aligned} L[X_i](s) &= \int_0^\infty X_i(t) e^{-st} dt, \\ \mathcal{F}[X_i](\omega) &= \int_{-\infty}^\infty X_i(t) e^{-j\omega t} dt, \end{aligned} \tag{1}$$

where $j = \sqrt{-1}$. The convolution between two functions f and g is defined as $f * g(t) := \int_{\mathbb{R}} f(x) g(t-x) dx$. The joint distribution of processes $\{X_1^n, ..., X_m^n\}$ is represented by $P_{\underline{X}}(n)$.

2.1 MULTIVARIATE HAWKES PROCESSES

Fix a complete probability space $(\Omega, \mathcal{F}, P)$. Let $N(t)$ denotes the counting process representing the cumulative number of events up to time t and let $\{\mathcal{F}^t\}_{t\geq 0}$ be a set of increasing σ-algebras such that $\mathcal{F}^t = \sigma\{N^t\}$. The non-negative, $\mathcal{F}^t$-measurable process $\lambda(t)$ is called the intensity of $N(t)$ if

$$P(N(t+dt) - N(t) = 1 | \mathcal{F}^t) = \lambda(t)dt + o(dt).$$

A classical example of mutually exciting processes, a multivariate Hawkes process [8], is a multidimensional process $\underline{N}(t) = \{N_1, ..., N_m\}$ such that for each $i \in [m]$

$$\begin{aligned} &P\left(dN_i(t) = 1 | \underline{\mathcal{F}}^t\right) = \lambda_i(t)dt + o(dt), \\ &P(dN_i(t) > 1 | \underline{\mathcal{F}}^t) = o(dt), \end{aligned} \tag{2}$$

where $\underline{\mathcal{F}}^t = \sigma\{\underline{N}^t\}$. The above equations imply that $\mathbb{E}[dN_i(t)/dt | \underline{\mathcal{F}}^t] = \lambda_i(t)$. Furthermore, the intensities are all positive and are given by

$$\lambda_i(t) = v_i + \sum_{k=1}^m \int_0^t \gamma_{i,k}(t-t') dN_k(t'). \tag{3}$$

The exciting functions $\gamma_{i,k}(\cdot)$s are in ℓ_1 such that $\lambda_i(t) \geq 0$ for all $t > 0$. Equivalently, in matrix representation:

$$\Lambda(t) = \mathbf{v} + \int_0^t \Gamma(t-t') d\underline{N}(t'), \tag{4}$$

where $\Gamma(\cdot)$ denotes an $m \times m$ matrix with entries $\gamma_{i,j}(\cdot)$; $d\underline{N}, \Lambda(\cdot)$, and $\mathbf{v}$ are $m \times 1$ arrays with entries $dN_i, \lambda_i(\cdot)$, and v_i, respectively. Matrix $\Gamma(\cdot)$ is called the excitation (kernel) matrix. Figure 1 illustrates the intensities of a multivariate Hawkes process comprised of two processes ($m = 2$) with the following parameters

$$\mathbf{v} = \begin{pmatrix} 0.5 \\ 0.4 \end{pmatrix}, \quad \Gamma(t) = \begin{pmatrix} 0.1e^{-t} & 0.3e^{-1.1t} \\ 0.5e^{-0.9t} & 0.3e^{-t} \end{pmatrix} u(t),$$

where $u(t)$ is the unit step function.

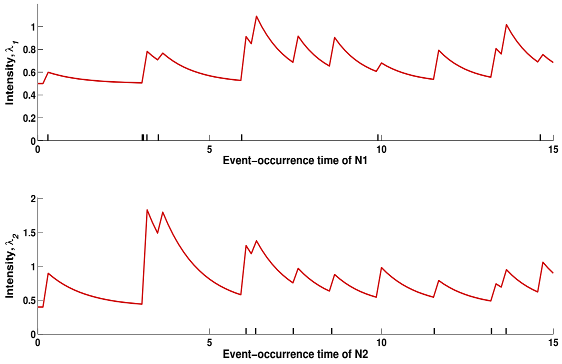


Figure 1: Intensities of the multivariate Hawkes process.

Assumption 1 *A joint distribution is called positive (non-degenerate), if there exists a reference measure ϕ such that $P_{\underline{X}} \ll \phi$ and $\frac{dP_{\underline{X}}}{d\phi} > 0$, where $P_{\underline{X}} \ll \phi$ denotes that $P_{\underline{X}}$ is absolutely continuous with respect to ϕ*[1].

Note that the Assumption 1 states that none of the processes is fully determined by the other processes.

2.2 CAUSAL STRUCTURE

A causal model allows the factorization of the joint distribution in some specific ways. *Generative model graphs* are a type of graphical model that similar to Bayesian networks [17] represent a causal factorization of the joint [19]. More precisely, it was shown in [19] that under Assumption 1, the joint distribution of a causal[2] discrete-time dynamical system with m processes can be factorized as follows,

$$P_{\underline{X}} = \prod_{i=1}^{m} P_{X_i||\underline{X}_{B_i}}, \qquad (5)$$

where $B(i) \subseteq -\{i\}$ is the minimal[3] set of processes that causes process X_i, i.e., parent set of node i in the corresponding minimal generative model graph. Such factorization of the joint distribution is called minimal generative model. In Equation (5),

$$P_{X_i||\underline{X}_{B_i}} := \prod_{t=1}^{n} P_{X_i(t)|\underline{\mathcal{F}}^{t-1}_{B\cup\{i\}}},$$

and $\underline{\mathcal{F}}^{t-1}_{B\cup\{i\}} = \sigma\{\underline{X}^{t-1}_{B\cup\{i\}}\}$.

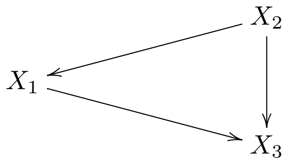


Figure 2: Minimal generative model graph of Example 1.

Extending the definition of generative model graphs to continuous-time systems requires some technicalities which are not necessary for the purpose of this paper. Hence we illustrate the general idea through an example.

The following example demonstrates the minimal generative model graph of a simple continuous-time system.

Example 1 *Consider a dynamical system in which the processes evolve over time horizon $[0, T]$ through the following coupled differential equations:*

$$dX_1 = f(X_1, X_2)dt + dW,$$
$$dX_2 = g(X_2)dt + dU,$$
$$dX_3 = h(X_1, X_2, X_3)dt + dV,$$

where W, U and V are independent exogenous noises. For small time dt, this becomes,

$$\begin{aligned} dX_1(t+dt) &\approx \Delta f(X_1(t), X_2(t)) + dW(t),\\ dX_2(t+dt) &\approx \Delta g(X_2(t)) + dU(t), \qquad (6)\\ dX_3(t+dt) &\approx \Delta h(X_1(t), X_2(t), X_3(t)) + dV(t). \end{aligned}$$

In this example, since the system is causal, the corresponding joint distribution can be factorized as follows,

$$P_{\underline{X}} = \prod_{j=1}^{3} \prod_{k\geq 0} P_{X_j(T-kdt)|\underline{\mathcal{F}}^{T-(k+1)dt}}, \qquad (7)$$

where $\underline{\mathcal{F}}^{T-(k+1)dt} = \sigma\{\underline{X}^{T-(k+1)dt}_{\{1,2,3\}}\}$. Due to (6), we can rewrite (7) as

$$P_{\underline{X}} = P_{X_1||X_2} P_{X_2} P_{X_3||X_1,X_2}. \qquad (8)$$

Figure 2 demonstrates the corresponding generative model graph of the factorization in (8).

In general, the joint distribution of a causal dynamical system can be factorized as $P_{\underline{X}} = \prod_{i=1}^{m} P_{X_i||\underline{X}_{B_i}}$, where $B(i) \subseteq -\{i\}$ is the parent set of node i in the corresponding minimal generative model graph, and

$$P_{X_i||\underline{X}_{B_i}} = \prod_{k\geq 0} P_{X_i(T-kdt)|\underline{\mathcal{F}}^{T-(k+1)dt}_{B_i}}.$$

3 TWO EQUIVALENT NOTATIONS OF CAUSALITY FOR HAWKES PROCESSES

In linear multivariate Hawkes processes, a natural notion of causation exists in the following sense: if $\gamma_{i,j} \neq 0$, then

[1] A measure $P_{\underline{X}}$ on Borel subsets of the real line is absolutely continuous with respect to measure ϕ if for every measurable set B, $\phi(B) = 0$ implies $P_{\underline{X}}(B) = 0$.

[2] In causal systems, given the full past of the system, the present of the processes become independent.

[3] Minimal in terms of its cardinality.

occurrence of an event in jth process will affect the likelihood of the arrivals in ith process. Next, we establish the relationship between the excitation matrix of multivariate Hawkes processes and their generative model graph. To do so, first, we discuss the equivalence of directed information graphs and generative models graphs which was established in [20].

3.1 DIRECTED INFORMATION GRAPHS (DIGs)

An alternative graphical model to encode statistical interdependencies in stochastic causal dynamical systems are *directed information graphs* (DIGs) [19]. Such graphs are defined based on an information-theoretic quantity, *directed information* (DI) that generalizes the Granger causality and it was shown in [20] that under some mild assumptions, they are equivalent to the minimal generative model graphs. Hence, DIGs also represent a minimal factorization of the joint distribution.

In a DIG, to determine whether X_j causes X_i over a time horizon $[0, T]$ in a network of m random processes, two conditional probabilities are compared in KL-divergence sense: one is the conditional probability of $X_i(t+dt)$ given full past, i.e., $\underline{\mathcal{F}}^t := \sigma\{\underline{X}^t\}$ and the other one is the conditional probability of $X_i(t+dt)$ given full past except the past of X_j, i.e., $\underline{\mathcal{F}}^t_{-\{j\}} := \sigma\{\underline{X}^t_{-\{j\}}\}$. It is declared that there is no influence from X_j on X_i, if the two conditional probabilities are the same. More precisely, there is an influence from X_j on X_i if and only if the following directed information measure is positive [19],

$$I_T(X_j \to X_i || \underline{X}_{-\{i,j\}}) := \inf_{\mathbf{t} \in \mathcal{T}(0,T)} \tilde{I}_{\mathbf{t}}(X_j \to X_i || \underline{X}_{-\{i,j\}}), \tag{9}$$

where $-\{i, j\} := [m] \setminus \{i, j\}$, $\mathcal{T}$ denotes the set of all finite partitions of the time interval $[0, T]$ [23], and

$$\tilde{I}_{\mathbf{t}}(X_j \to X_i || \underline{X}_{-\{i,j\}}) := \sum_{k=0}^{n} I\left(X^{t_k}_{i,t_{k-1}}; X^{t_k}_{j,0} | \mathcal{F}^{t_{k-1}}_{-\{j\}} \right),$$

where $\mathbf{t} := (0 = t_0, t_1, ..., t_n = T)$. Finally, $I(X; Y|Z)$ represents the conditional mutual information between X and X given Z and it is given by

$$I(X; Y|Z) := \mathbb{E}_{P_{X,Y,Z}} \left[\log \frac{dP_{X|Y,Z}}{dP_{X|Z}} \right].$$

3.2 EQUIVALENCE BETWEEN GENERATIVE MODEL GRAPHS AND SUPPORT OF EXCITATION MATRIX

As mentioned earlier, the corresponding minimal generative model graph and the DIG of a causal dynamical system are equivalent. Thus, to characterize the corresponding minimal generative model graphs of a multivariate Hawkes system, we study the properties of its corresponding DIG.

Proposition 1 *Consider a set of mutually exciting processes $\underline{N}$ with excitation matrix $\Gamma(t)$. Under Assumption 1, $I_T(N_j \to N_i || \underline{N}_{-\{i,j\}}) = 0$ if and only if $\gamma_{i,j} \equiv 0$ over time interval $[0, T]$.*

Proof: See Section 7.1. □

Proposition 1 signifies that the support of the excitation matrix $\Gamma(\cdot)$ determines the adjacency matrix of the DIG and vice versa. Therefore, learning DIG of a mutually exciting Hawkes processes satisfying Assumption 1 is equivalent to learning the excitation matrix given samples from each of the processes. In other word, in the presence of side information that the processes are Hawkes, it is more efficient to learn the causal structure through learning the excitation matrix rather than the directed information needed for learning the DIG in general.

4 LEARNING THE EXCIATIONA MATRIX

In this section, we present an approach for learning the causal structure of a stationary Hawkes network with exponential exciting functions through learning the excitation matrix. This method is based on second order statistic of the Hawkes processes and it is suitable for the case when no i.i.d. samples are available. Note that when i.i.d. samples are available, non-parametric methods for learning the excitation matrix such as MMEL algorithm [25] exist. In this approach the exciting functions are expressed as linear combination of a set of base kernels and a penalized likelihood is used to estimate the parameters of the model. As mentioned earlier, we focus on learning the excitation matrix of multivariate Hawkes processes with exponential exciting functions. This class of Hawkes processes has been widely applied in many areas such as seismology, criminology, and finance [15, 21, 13, 5].

Definition 2 *The excitation matrix of a multivariate Hawkes processes with exponential exciting functions is defined as follows*

$$\mathcal{E}_{xp}(m) := \{\sum_{d=1}^{D} A_d e^{-\beta_d t} u(t) : A_d \in \mathbb{R}^{m \times m},$$
$$(\sum_{d=1}^{D} A_d e^{-\beta_d t})_{i,j} \geq 0, \rho(\sum_{d=1}^{D} \frac{A_d}{\beta_d}) < 1, D \in \mathbb{N}\}, \tag{10}$$

where $\{\beta_d\} > 0$ is called the set of exciting modes.

Example 2 *Consider a set of $m = 5$ mutually exciting processes with the following exponential excitation matrix*

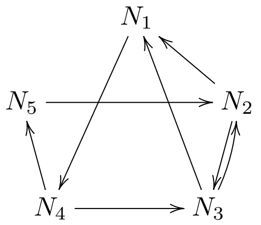


Figure 3: Corresponding DIG of the network in Example 2 with the excitation matrix given by (11)

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & 0 \\ 0 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \frac{e^{-t}}{20} + \begin{pmatrix} 0 & 0 & .5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2.5 & 0 \\ .1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \frac{e^{-1.4t}}{20}$$
$$+ \begin{pmatrix} 1 & 1.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \frac{e^{-2t}}{20} \qquad (11)$$

In this example $D = 3$ and the exciting modes are $\{1, 1.4, 2\}$. By Proposition 1, the adjacency matrix of the corresponding DIG of this network is given by the support of its excitation matrix. Figure 3 depicts the corresponding DIG.

Before describing our algorithm, we need to derive some useful properties of moments of the process. A multivariate Hawkes process with the excitation matrix Γ has stationary increments, i.e., the intensity processes is stationary, if and only if the following assumption holds [8, 6]:

Assumption 2 *The spectral radius (the supremum of the absolute values of the eigenvalues) of the matrix $\overline{\Gamma}$, where $[\overline{\Gamma}]_{i,j} = ||\gamma_{i,j}||_1$ is strictly less than one, i.e., $\rho(\overline{\Gamma}) < 1$.*

In this case, from (4) and Equation (2):

$$\Lambda = \mathbb{E}[\Lambda(t)] = \mathbf{v} + \int_0^t \Gamma(t-t')\mathbb{E}[d\underline{N}(t')]$$
$$= \mathbf{v} + \int_0^t \Gamma(t-t')\Lambda dt' = \mathbf{v} + \overline{\Gamma}\Lambda. \qquad (12)$$

By Assumption 2, $\sum_{i\geq 0} \overline{\Gamma}^i$ converges to $(I-\overline{\Gamma})^{-1}$, thus $\Lambda = (I-\overline{\Gamma})^{-1}\mathbf{v}$. The normalized covariance matrix of a stationary multivariate Hawkes process with lag τ and window size $z > 0$ is defined by

$$\Sigma_z(\tau) := \frac{1}{z}\mathbb{E}\left[\int_t^{t+z} d\underline{N}(x) \int_{t+\tau}^{t+\tau+z} (d\underline{N}(y))^T\right] - \Lambda\Lambda^T z, \qquad (13)$$

where $\int_t^{t+t'} d\underline{N}(x)$ denotes the number of events in time interval $(t, t+t']$.

Theorem 3 *[1] The Fourier transform of the normalized covariance matrix of a stationary multivariate Hawkes process with lag τ and window size $z > 0$ is given by*

$$\mathcal{F}[\Sigma_z](-\omega) \qquad (14)$$
$$= 4\frac{\sin^2 z\omega/2}{\omega^2 z}\left(I - \mathcal{F}[\Gamma](\omega)\right)^{-1} diag(\Lambda) \left(I - \mathcal{F}[\Gamma](\omega)\right)^{-\dagger},$$

where $A^\dagger$ denotes the Hermitian conjugate of matrix A, and $diag(\Lambda)$ is a diagonal matrix with vector Λ as the main diagonal.

In order to learn the excitation matrix with exponential exciting functions, we need to learn the exciting modes $\{\beta_d\}$, the number of components D, and coefficient matrices $\{A_d\}$. Next results establishes the relationship between the exciting modes and the number of components D with the normalized covariance matrix of the process.

Corollary 4 *Consider a network of a stationary multivariate Hawkes processes with excitation matrix $\Gamma(t)$ belonging to $\mathcal{E}xp(m)$. Then the exciting modes of $\Gamma(t)$ are the absolute values of the zeros of $1/\operatorname{Tr}\mathcal{F}[\Sigma_z]^{-1}(\omega)$.*

Proof: See Section 7.2. □

Next, we need to find the coefficient matrices $\{A_d\}$. To do so, we use the covariance density of the processes. The covariance density of a stationary multivariate Hawkes process for $\tau > 0$ is defined as [8]

$$\Omega(\tau) := \mathbb{E}\left[(d\underline{N}(t+\tau)/dt - \Lambda)(d\underline{N}(t)/dt - \Lambda)^T\right]. \qquad (15)$$

Since the processes have stationary increments, we have $\Omega(-\tau) = \Omega^T(\tau)$.

Lemma 5 *[8]*

$$\Omega(\tau) = \Gamma(\tau)diag(\Lambda) + \Gamma * \Omega(\tau), \tau > 0. \qquad (16)$$

It has been shown in [3] that the above equation admit a unique solution for $\Gamma(\tau)$. Next proposition provides a system of linear equations that allows us to learn the coefficient matrices.

Proposition 6 *Consider a network of a stationary multivariate Hawkes processes with excitation matrix $\Gamma(t) \in \mathcal{E}xp(m)$, and exciting modes $\{\beta_1, ..., \beta_D\}$. Then $\{A_d\}$ are a solution of the linear system of equations: $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{H}$, where $\boldsymbol{H}_{m^2\times m^2}$ is a block matrix with (i,j)th block given by*

$$\boldsymbol{H}_{i,j} = \frac{diag(\Lambda) + \mathcal{L}[\Omega](\beta_j) + \mathcal{L}[\Omega]^T(\beta_i)}{\beta_j + \beta_i},$$

and $\boldsymbol{A} = [A_1, ..., A_D]$ and $\boldsymbol{S} = [\mathcal{L}[\Omega](\beta_1), ..., \mathcal{L}[\Omega](\beta_D)]$.

Proof: See Section 7.3.□

Combining the results of Corollary 4 and Proposition 6 allows us to learn the excitation matrix of exponential multivariate Hawkes processes from the second order moments. Consequently applying Proposition 1, the causal structure of the network can be learned by drawing an arrow from node i to j, when $\sum_{d=1}^{D} |(A_d)_{j,i}| > 0$.

4.1 ESTIMATION AND ALGORITHM

This section discusses estimators for the second order moments, namely the normalized covariance matrix and the covariance density of a stationary multivariate Hawkes processes from data. Once such estimators are available, the approach of previous section maybe used to learn the network. The most intuitive estimator for Λ defined by Equation (12) is $\underline{N}(T)/T$. It turns out that this estimator converges almost surely to Λ as T goes to infinity [2]. Furthermore, [2] proposes an empirical estimator for the normalized covariance matrix as follows

$$\widehat{\Sigma}_{z,T}(\tau) := \frac{1}{T} \sum_{i=1}^{\lfloor T/z \rfloor} (X_{iz} - X_{(i-1)z})(X_{iz+\tau} - X_{(i-1)z+\tau})^T, \tag{17}$$

where $X_t := \underline{N}(t) - \Lambda t$. In the same paper, it has been shown that under Assumption 2, the above estimator converges in ℓ_2 to the normalized covariance matrix (13), i.e., $\widehat{\Sigma}_{z,T}(\tau) \underset{T\to\infty}{\longrightarrow} \Sigma_z(\tau)$. Notice that the normalized covariance matrix and the covariance density are related by $\Sigma_{dt}(\tau)/dt = \Omega^T(\tau)$. Therefore, we can estimate the covariance density matrix using Equation (17) by choosing small enough window size $z = \Delta$. Namely, $\widehat{\Omega}_{\Delta}^T(\tau) = \widehat{\Sigma}_{\Delta}(\tau)/\Delta$.

Algorithm 1

```
$Input:\ \underline{N}^T$.
$Output:$ DIG.
$\widehat{\Lambda} \leftarrow \underline{N}(T)/T$
Choose $\sigma > 0$, $z > 0$, and small $\Delta > 0$.
Compute $\widehat{\Sigma}_{z,T}(\tau)$ and $\widehat{\Omega}_{\Delta}(\tau)$ using (17).
$\{\widehat{\beta}_d\}_{d=1}^{\widehat{D}} \leftarrow$ Zeros of $1/\operatorname{Tr}\mathcal{F}[\Sigma_z]^{-1}(\omega)$.
Compute $\mathcal{L}[\widehat{\Omega}_{\Delta}](\widehat{\beta}_d)$ for $d = 1, ..., \widehat{D}$.
Solve the set of equations arises from (20) for $\widehat{A}_d$.
Draw $(j,i)$ if $\sum_{d=1}^{\widehat{D}} |(\widehat{A}_d)_{i,j}| \geq \sigma$.
```

Algorithm 1 summarizes the steps of our proposed approach for learning the excitation matrix and consequently the causal structure of an exponential multivariate Hawkes process.

5 EXPERIMENTAL RESULTS

In this section, we present our experimental results for both synthetic and real data.

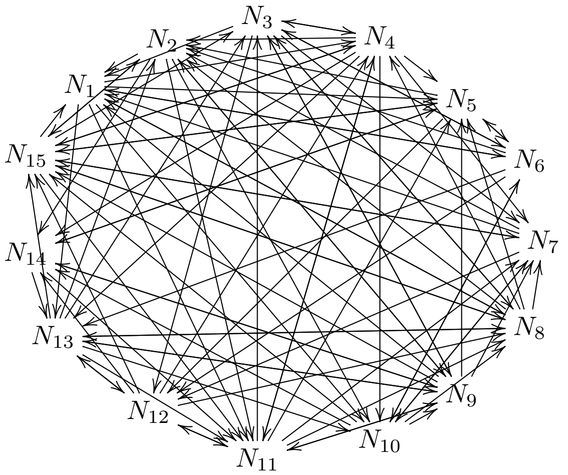


Figure 5: True causal structure of the synthesized example.

5.1 SYNTHETIC DATA

We apply the proposed algorithms to learn the causal structure of the multivariate Hawkes network of Example 2 with $\mathbf{v} = (0.5, 0.4, 0.5, 1, 0.3)^T$. This network satisfies Assumption 2, since $\rho(\overline{\Gamma}) \approx 0.16$. The exciting modes are $\{1, 1.4, 2\}$. We observed the arrivals of all processes during a time period T. Figure 4 depicts the outputs of algorithms 1 for $\Delta = 0.2$, $z = 2$, and observation lengths $T \in \{1000, 2100\}$. As illustrated in Figure 4, by increasing the length of observation T, the output graph converges the true DIG shown in Figure 3. As a comparison, we applied the MMEL algorithm proposed in [25] to learn the excitation matrix for this example and the numerical method based on Nystrom method proposed in [3] with $T = 2100$ and the number of quadrature $Q = 70$. Since MMEL requires i.i.d. samples, we generate 35 i.i.d. samples each of length 60 to obtain Figure 4(MMEL). Our proposed algorithm outperforms both MMEL and the numerical method of [3].

Furthermore, we conducted another experiment for a network of 15 processes with 102 edges illustrated in Figure 5. For a sample of length $T = 2500$, our algorithm was able to recover 70 edges correctly but identified 34 false arrows. MMEL could only recover 58 arrows correctly while detecting another 41 false arrows. The input for MMEL was 25 sequences each of length 100.

5.2 STOCK MARKET DATA

As an example of how our approach may discover causal structure in real-world data, we analyzed the causal relationship between stock prices of 12 technology companies of the New York Stock Exchange sourced from Google Finance. The prices were sampled every 2 minutes for twenty market days (03/03/2008 - 03/28/2008). Every time a stock price changed by ±1% of its current price an event was

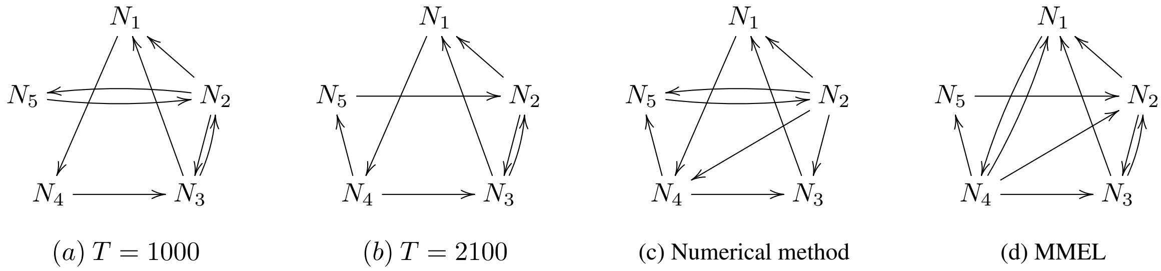
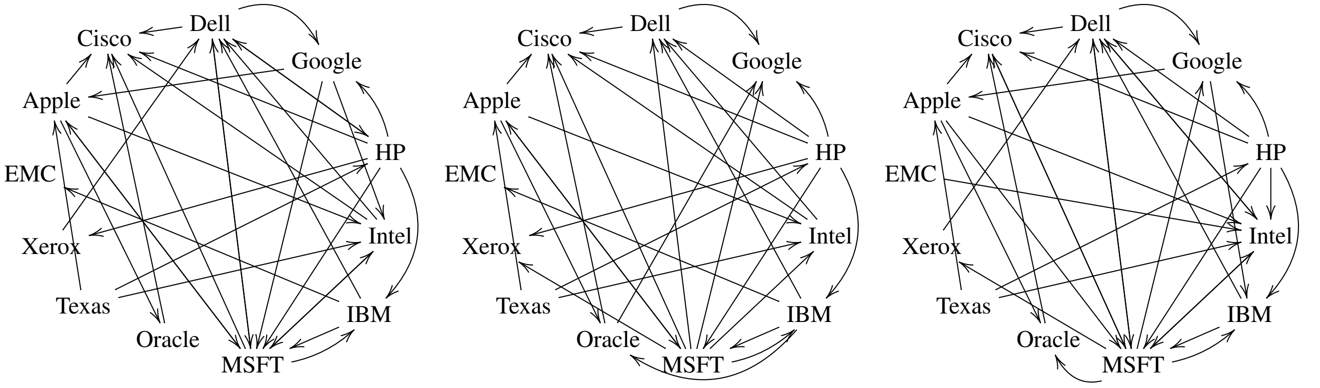


Figure 4: Recovered DIG of the network in Example 2 with the excitation matrix given by (11), (a), (b) Algorithm 1 with $\Delta = 0.2$, $z = 2$, and $T \in \{1000, 2100\}$, (c) the numerical method of [3] with $Q = 70$ and $T = 2100$, and (d) MMEL with 35 i.i.d. samples each of length 60. Our approach learns the graph with $T = 2100$, while other approaches fail at the same sample size.



(a) Algorithm 1.

(b) DIG.

(c) MMEL.

Figure 6: Causal structures for the S&P (a) using Algorithm 1, (b) by estimating the directed information DIG, and (c) using MMEL algorithm.

logged on the stock's process. In order to prevent the substantial changes in stock's prices due to the opening and closing of the market, we ignored the samples at the beginning and at the end of each working day. For this part, we have assumed that the jumps occurring in stock's prices are correlated through a multivariate Hawkes process. This model class was advocated in [11, 2]. Figure 6(a) illustrate the causal graph resulting from Algorithm 1, with $z = 30$ and $\Delta = 2$ minutes.

To compare our learning approach with other approaches, we applied the MMEL algorithm to learn the corresponding causal graph. For this scenario, we assumed that the data collected from each day is generated i.i.d. Hence, a total of 20 i.i.d. samples were used. Figure 6(c) illustrates the resulting graph. As one can see, Figures 6(a) and 6(c) convey pretty much a similar causal interactions in the dataset. For instance both of these graphs suggest that one of the most influential companies in that period of time was Hewlett-Packard (HP). Looking into the global PC market share during 2008, we find that this was indeed the case.[4]

To use another modality, we derive the corresponding DIG of this network applying Equation (9). For this part, we used the market based on the Black-Scholes model [4] in which the stock's prices are modeled via a set of coupled stochastic PDEs. We assumed that the logarithm of the stock's prices are jointly Gaussian and therefore the corresponding DIs were estimated using Equation (24) in [7]. The resulting DIG is shown in Figure 6(b). Note that this DIG is derived from the logarithm of prices and not the jump processes we used earlier. Still it shares a lot of similarities with the two other graphs. For instance, it also identifies HP as one of the most influential companies and Microsoft as one the most influenced companies in that time period.

	Alg. 1	DIG	MMEL
Alg. 1	33	25	26
DIG	25	30	24
MMEL	26	24	34

This table shows the number of edges that each of the above approaches recovers and the number of edges that they jointly recover. This demonstrates the power of exponential kernels even when data does not come from such a model class.

[4] Gartner, http://www.gartner.com/newsroom/id/856712

5.3 MEMETRACKER DATA

We also studied causal influences in a blogosphere. The causal flow of information between media sites may be captured by studying hyperlinks provided in one media site to others. Specifically, the time of such linking can be modeled using a linear multivariate Hawkes processes with exponential exciting functions [25, 18]. This model is also intuitive in the sense that after emerging a new hot topic, in the first several days, the blogs or websites are more likely feature that topics and it is also more likely that the topic would trigger further discussions and create more hyperlinks. Thus, exponential exciting functions are well suited to capture such phenomenon as the exiting functions should have relatively large values at first and decay fast as time elapses.

For this experiment, we used the MemeTracker[5] dataset. The data contains time-stamped phrase and hyperlink information for news media articles and blog posts from over a million different websites. We extracted the times that hyperlinks to 10 well-known websites listed in Table 1 are created during August 2008 to April 2009. When a hyperlink to a website is created at a certain time, an arrival events is recorded at that time. More precisely, in this experiment, we picked 30 different phrases that appeared on different websites at different times. If a website that published one of the phrases at time t also contained a hyperlink to one of the 10 listed websites, an arrival event was recorded at time t for that website in our list.

Figure 7(a) illustrates the resulting causal structure learned by Algorithm 1 for $z = 12$ hours and $\Delta = 1$ hour. In this graph, an arrow from a node to another, say node Ye to Yo, means creating a hyperlink to `yelp.com` triggers creation of further hyperlinks to `youtube.com`.

We also applied the MMEL algorithm with one exponential kernel function to learn the excitation matrix. For this experiment, the data corresponding to each phrase was treated as an i.i.d. realization of the system. The resulting causal structure is depicted in Figure 7(b).

As Figure 7(a) illustrates, the nodes can be clustered into two main groups: $\{$Cr, Ye, Am, Yo$\}$ and $\{$Bb, Cn, Gu, Hu, Sp, Wi$\}$. The first group consists of mainly merchandise and reviewing websites and the second group contains the broadcasting websites. However, this is not as clear in Figure 7(b). This is because MMEL requires more i.i.d. samples (phrases) to be able to identify the correct arrows. Note that as we increase the number of phrases (110), Figure 7(c), both graphs become similar with two clearly visible main clusters.

Cr	`craigslist.org`
Ye	`yelp.com`
Am	`amazon.com`
Sp	`spiegel.de`
Wi	`wikipedia.org`
Yo	`youtube.com`
Cn	`cnn.com`
Gu	`guardian.co.uk`
Hu	`humanevents.com`
Bb	`bbc.co.uk`

Table 1: List of websites studied in MemeTracker experiment.

6 CONCLUSION

Learning the causal structure (DIG) of a stochastic network of processes requires estimation of conditional directed information (9). Estimating this quantity in general has high complexity and requires a large number of samples. However, the complexity of the learning task could be significantly reduced, if side information about the underlying structure of system dynamics is available. As proved in 1, for multivariate Hawkes processes, estimating the support of the excitation matrix suffices to learn the associated DIG. Therefore, all approaches for learning the excitation matrix of the multivariate Hawkes processes such as ML estimation [16, 25], EM algorithm [10], non-parametric estimation techniques proposed in [2], and the proposed method in this paper may be used to learn the causal interactions in such networks. The previous estimation approaches either require i.i.d. samples such as MMEL or are limited to the class of symmetric Hawkes processes. The proposed algorithm in this work allows us to learn the support of the excitation matrix in a larger class of matrices in the absence of i.i.d. samples.

Acknowledgements

This work was in part supported by NSF CCF 10- 54937 - CAREER, MURI grant ARMY W911NF-15-1-0479, and Army W911NF-15-1-0281 AB262.

7 TECHNICAL PROOFS

7.1 Proof of Proposition 1

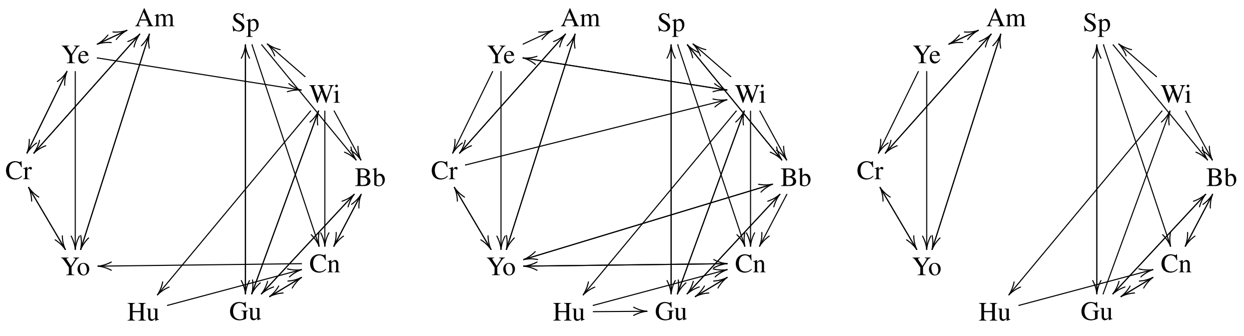
Suppose $\gamma_{i,j} \equiv 0$. (3) implies that for every $t \leq T$, $\lambda_i(t)$ is $\underline{\mathcal{F}}^t_{-\{j\}}(= \sigma\{\underline{N}^t_{-\{j\}}\})$-measurable and from (2), we have

$$P\left(dN_i(t) = 1|\underline{\mathcal{F}}^t\right) = P(dN_i(t) = 1|\underline{\mathcal{F}}^t_{-\{j\}}).$$

Equivalently, for every $0 \leq t_{k-1} < t_k$,

$$I\left(N^{t_k}_{i,t_{k-1}}; N^{t_k}_{j,0}|\mathcal{F}^{t_{k-1}}_{-\{j\}}\right) = 0, \tag{18}$$

[5] `http://memetracker.org/data/links.html`



(a) Alg. 1 (30).

(b) MMEL.

(c) Both Alg. 1 & MMEL (110).

Figure 7: Recovered causal structure of the MemeTracker dataset using (a) Algorithm 1, (b) MMEL for 30 different phrases, and (c) both Algorithm 1 and MMEL for 110 different phrases.

and thus, $\tilde{I}_{\mathbf{t}}(N_j \to N_i||\underline{N}_{-\{i,j\}}) = 0$, for any finite partition $\mathbf{t} \in \mathcal{T}(0,T)$.

For the converse we use proof by contradiction. Suppose $I_T(N_j \to N_i||\underline{N}_{-\{i,j\}}) = 0$ and $\gamma_{i,j} \neq 0$. Using the definition in (9), it is straightforward to observe that for any $t < T$,

$$I_t(N_j \to N_i||\underline{N}_{-\{i,j\}}) = 0.$$

Similarly, $I_{t+dt}(N_j \to N_i||\underline{N}_{-\{i,j\}}) = 0$. Consequently,

$$\begin{aligned} 0 &= I_{t+dt}(N_j \to N_i||\underline{N}_{-\{i,j\}}) - I_t(N_j \to N_i||\underline{N}_{-\{i,j\}}) \\ &= I\left(dN_i(t); N_{j,0}^t|\mathcal{F}_{-\{j\}}^t\right). \end{aligned}$$

This implies $P(dN_i(t) = 1|\underline{\mathcal{F}}_{-\{j\}}^t) = \lambda_i(t)dt + o(dt)$, or $\lambda_i(t)$ is $\underline{\mathcal{F}}_{-\{j\}}^t$-measurable. Since, we have assumed $\gamma_{i.j} \neq 0$, we obtain $N_j(t)$ is $\underline{\mathcal{F}}_{-\{j\}}^t$-measurable, for all $t \leq T$. In words, jth process is determined by other processes which contradicts with the Assumption 1 that states there is no deterministic relationships between processes.

7.2 Proof of Corollary 4

If the excitation matrix belongs to $\mathcal{E}xp(m)$, from Equation (14) we have

$$\left(I - \sum_{d=1}^{D} \frac{A_d^T}{j\omega + \beta_d}\right) diag(\Lambda)^{-1} \left(I - \sum_{d=1}^{D} \frac{A_d}{-j\omega + \beta_d}\right) = \frac{4\sin^2 z\omega/2}{\omega^2 z}\mathcal{F}[\Sigma_z]^{-1}(\omega).$$

By evaluating the trace of the above equation, we obtain

$$\sum_{i=1}^{m} \frac{|1-a_{i,i}|^2}{\lambda_i} + \sum_{i\neq j} \frac{|a_{i,j}|^2}{\lambda_i} = \frac{4\sin^2 z\omega/2}{\omega^2 z} \operatorname{Tr} \mathcal{F}[\Sigma_z]^{-1}(\omega), \tag{19}$$

where $a_{i,j} = \sum_{d=1}^{D} \frac{a_{i,j}^{(d)}}{-j\omega+\beta_d}$, and $A_d = [a_{i,j}^{(d)}]$. To learn the entire set $\{\pm j\beta_d\}$, we have to show that there are no pole zero cancellations in (19). That is, the nominator and denominator of (19) have no common roots. Let

$$g(\omega) := \left(\sum_{i=1}^{m} \frac{|1-a_{i,i}|^2}{\lambda_i} + \sum_{i\neq j} \frac{|a_{i,j}|^2}{\lambda_i}\right) \prod_{d=1}^{D} |{-j\omega}+\beta_d|^2,$$

which is the nominator of Equation (19). It is straightforward to check that for $\omega = -j\beta_k$, the above quantity is non-zero, due to the fact that β_ds are distinct and $A_k \neq \mathbf{0}$. Since $g(\omega)$ is a polynomial with real coefficients, from complex conjugate root theorem [9], we have $g(j\beta_k) \neq 0$. Therefore, the set $\{\pm j\beta_d\}$ contains all the poles of (19).

7.3 Proof of Proposition 6

From Lemma 5, the Laplace transform of the covariance density can be written as

$$\begin{aligned} \mathcal{L}[\Omega](s) &= \mathcal{L}[\Gamma](s)\left(diag(\Lambda) + \mathcal{L}[\Omega](s)\right) \\ &+ \int_0^\infty \int_t^\infty \Gamma(t')\Omega^T(t)e^{-s(t'-t)}dt'dt. \end{aligned}$$

When $\Gamma(t) \in \mathcal{E}xp(m)$, it can be shown that (1) becomes

$$\mathcal{L}[\Omega](s) = \sum_{d=1}^{D} \frac{A_d}{s+\beta_d}\left(diag(\Lambda) + \mathcal{L}[\Omega](s) + \mathcal{L}[\Omega]^T(\beta_d)\right). \tag{20}$$

If the set of exciting modes are given, we can insert $s = \beta_d$, for $d = 1, \ldots, D$ in the above equation and obtain the system of D equations.

References

[1] Emmanuel Bacry, Khalil Dayri, and Jean-Francois Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12, 2012.

[2] Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-Francois Muzy. Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499, 2013.

[3] Emmanuel Bacry and Jean-Francois Muzy. Second order statistics characterization of hawkes processes and non-parametric estimation. *preprint arXiv:1401.0903*, 2014.

[4] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

[5] Clive G Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[6] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

[7] Jalal Etesami and Negar Kiyavash. Directed information graphs: A generalization of linear dynamical graphs. In *American Control Conference (ACC), 2014*, pages 2563–2568. IEEE.

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

[9] Alan Jeffrey. *Complex analysis and applications*, volume 10. CRC Press, 2005.

[10] Erik Lewis and George Mohler. A nonparametric em algorithm for multiscale hawkes processes. *Preprint*, 2011.

[11] Scott W Linderman and Ryan P Adams. Discovering latent network structure in point process data. *preprint arXiv:1402.0914*, 2014.

[12] Thomas Josef Liniger. *Multivariate hawkes processes*. PhD thesis, Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403, 2009, 2009.

[13] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011.

[14] Ioane Muni Toke and Fabrizio Pomponio. Modelling trades-through in a limited order book using hawkes processes. *Economics discussion paper*, (2011-32), 2011.

[15] Yosihiko Ogata. Seismicity analysis through point-process modeling: A review. *Pure and Applied Geophysics*, 155(2-4):471–507, 1999.

[16] T Ozaki. Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

[17] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[18] Julio Cesar Louzada Pinto, Tijani Chahed, and Eitan 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.

[19] Christopher Quinn, Negar Kiyavash, and Todd P Coleman. Directed information graphs. *Transactions on Information Theory*, 61(12):6887–6909, 2015.

[20] Christopher J Quinn, Negar Kiyavash, and Todd P Coleman. Equivalence between minimal generative model graphs and directed information graphs. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 293–297. IEEE, 2011.

[21] Patricia Reynaud-Bouret, Sophie Schbath, et al. Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

[22] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[23] Tsachy Weissman, Young-Han Kim, and Haim H Permuter. Directed information, causal estimation, and communication in continuous time. *Information Theory, IEEE Transactions on*, 59(3):1271–1287, 2013.

[24] Shuang-Hong Yang and Hongyuan 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.

[25] Ke Zhou, Hongyuan Zha, and Le 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.