Discrete Time Latent Hawkes Processes for Modeling Multidimensional Temporal Event Streams

Anonymous Author
Anonymous Institution

Abstract

Multidimensional event streams are common in many applications such as content on social platforms. Existing models using Hawkes processes and its variants often ignore important information about the causal parents of the events, which typically is readily available in social media applications. These models also ignore the disproportionate response created by some of the rare events, such as the impact of a "like" on a content by an influencer. Addressing these limitations, this paper proposes a novel Bayesian dIscRete Time Hawkes (BIRTH) model, a Bayesian generative model for multidimensional event streams data with causal information. Through its latent variables, BIRTH is flexible enough to capture contrasting responses invoked by multiple events of the same type. Moreover, being a discrete-time model, the latent parameters scale as $O(\#Timebins)$ in BIRTH as compared to $O(\#events)$ for continuous-time processes, thus scaling better in the settings when the number of events is huge. For inference, we propose a Gibbs sampling based inference procedure for BIRTH, which is suitable when the whole data can be processed together. While a full variational inference procedure is difficult to arrive at due to non-conjugate factors in the posterior, we propose a Stochastic hybrid Gibbs-Variational Inference (SVI) algorithm, which is beneficial in the settings where Gibbs might be expensive in terms of memory requirements. SVI has per-iteration memory complexity proportional to the chosen minibatch size, and also extends easily for online streaming settings of the data. We thoroughly evaluate BIRTH's abilities over both synthetic and real-world social network event streams. Specifically, on synthetic datasets we demonstrate model fitting, recovery of planted structure and identification of the rare events. For a social network dataset we show significantly higher likelihoods along with rare event identification.

1 Introduction

Multi-dimensional temporal event streams are common in a variety of domains such as social networks [28, 23, 20], neuro-science [8], e-commerce [26], finance [10, 1], to name a few. Having observed part of a stream, modeling and predicting its future course helps in various ways. For instance, predicting which posts in a social media platform would go viral helps in reviewing them in a focused way leading to better quality of content served to the users. While many approaches focus more on engineering or synthesizing features to make use of prediction algorithms [24, 15, 6, 18, 14, 5], other approaches focus on modeling the evolution of the streams using generative approaches such as Hawkes Processes [9, 27, 20, 8, 17]. In this paper, we focus on modeling event streams while addressing the shortcomings of the existing methods. Our approach is generic and extend well beyond the application considered but for providing the reader with better clarity, we discuss a running example of modeling content streams on social media platforms throughout the paper.

A typical data sample observed in such streams is of the following form: $\mathcal{S} = \{(t_i, k_i) \,|t_i \in \mathbb{R}, k_i \in [K]\,, i \in [N]\}$, where N denotes the total number of events, K denotes the total number of event-types, the tuple (t_i, k_i) represents the i^{th} event where t_i and k_i denote the time and event type respectively. $\mathcal{S}$ is said to be a multi-dimensional event stream if $K > 1$. We make the following observations. First, in some of the domains, we also get to observe the causal parent for each event. For example, in the context of social networks where the event types constitute activities such as *Reshares, Likes, Comments, Views* on the shared content, typical events are in the following form: *Member A viewed the post x because one of A's connections, member B, has*

Preliminary work. Under review by AISTATS 2021. Do not distribute.

liked the post x. The post x, in turn, might have been shared by one of B's connections. The presence of these additional causal parent information is not typically exploited in many existing models. Second, the size of the data (N) in such applications is huge, which makes continuous-time techniques such as multivariate Hawkes processes [13, 20, 4] difficult to scale and they also ignore causal parent information. The best known technique [13] scales linearly in the number of events, which may still be expensive in scenarios such as viral event modeling. For example, in a social media setting, for viral streams the number of events can be in order of millions and typically all these events occur within a short duration (say one month). Continious time models have to process all these events for all the streams, hence they face a scalability issue. On the other hand, discretising these event into time bin (of one minute) brings down the scale from millions to thousands. This motivates one to consider discrete time variants of Hawkes processes [16]. Third, Hawkes models usually assume that the future influence of an event is determined by the event-type alone. While this may be modified to some extent by taking into account any marks or contextual features specific to the event [23], it does not yet completely capture the variability where two events of the same type and features may invoke completely different response for example in viral event modeling. In this paper, we make use of these observations and propose a novel graphical model addressing the limitations of the existing models. Specifically, we make the following contributions:

1. We propose BIRTH, a Bayesian Hawkes process with latent states for modeling multidimensional causal event streams (See Algorithm 1). In BIRTH, each event is mapped into a latent state through Multinomial sampling, which in turn captures the influence it has towards the future events. This removes the limitation that all events of the same type has similar influence throughout the process.

2. To estimate the model parameters and the latent variables of BIRTH, which has non-conjugate probability distributions, we first propose a Gibbs sampling based algorithm. This procedure utilizes complete data in each of the iterations and recovers model parameters with low variance, hence, useful when the data fits into memory.

3. Additionally, we propose a stochastic hybrid Gibbs-Variational inference (SVI) based procedure (see Algorithm 2), which works with minibatches of that data and thus has $\mathcal{O}(1)$ memory requirement for each of its iteration, making the procedure scalable for large datasets. This procedure naturally extends to an online settings (see Section 5.2) and uses gradient descent to update the global parameters of the model, while using Gibbs iterations to optimize the time-specific latent variables.

4. We illustrate the efficacy of the modeling approach on a real dataset consisting of viral event streams, which demonstrate that BIRTH comprehensively beats competetive baselines in terms of data likelihood, and also capture events which create disproportionate responses. Additionally, we also illustrate the efficacy of the algorithms proposed on simulated datasets.

2 Problem Setup

Over a finite time horizon divided into equally spaced time bins of length $\Delta > 0$, we consider a dataset $\mathcal{D}$ of M samples. The m^{th} data sample, denoted by $\mathcal{D}(m)$, is defined below:

$$\mathcal{D}(m) = \{\mathcal{D}_{t,k}(m) | k \in [K], t \in [T]\}\,, \text{where} \tag{1}$$
$$\mathcal{D}_{t,k}(m) = \left\{s^{(0)}_{t,k}(m)\right\} \bigcup \left\{ s^{t',k'}_{t,k}(m) \,\middle|\, k' \in [K], t' < t \right\}.$$

Here, $[x] = \{1, \ldots, x\}$, T denotes the maximum number of time bins, $\mathcal{D}_{t,k}(m)$ denote the observations made for the m^{th} data sample at time bin t towards an event type k. $\mathcal{D}_{t,k}(m)$ consists of the following: [1] (a) $s^{(0)}_{t,k}$, the total number of events of type k at bin t which are due to exogenous (external) factors, and (b) $s^{t',k'}_{t,k}$, the total number of events of type k at bin t which are caused by an event of type k' occurred in bin $t' < t$. Denoting by $\mathcal{E}_{t,k}(m)$, the set of all events of type k for the m^{th} sample at the time bin indexed by t, the quantities $s^{(0)}_{t,k}$ and $s^{t',k'}_{t,k}$ are formally defined as below:

$$s^{(0)}_{t,k}(m) = \mid \{e \in \mathcal{E}_{t,k}(m) \mid e \text{ is exogenous}\} \mid, \tag{2}$$
$$s^{t',k'}_{t,k}(m) = \mid \{e \in \mathcal{E}_{t,k}(m) \mid \exists e' \in \mathcal{E}_{t',k'}(m) \text{ s.t.}, e' \to e\} \mid, \tag{3}$$

where, $|.|$ is the cardinality of the set, $e \to e'$ denotes that the event e causes e'.

We assume that when Δ is sufficiently small, $s^{t',k'}_{t,k} = 0$ for $t' = t$. This is true in many applications such as social media, where the feed ranking system has inherent delays, thus a member A will be able to view an activity of his connection B only after the delay. As a convention throughout this paper, we use the common variable name s to denote counts of various quantities, where the superscripts denote the qualifiers for the source (bin, event-type, state, etc.), and subscripts denote the target bin, event-type. In the same notation, the total number of events of type k at time bin t, denoted by $s_{t,k}(m)$ equals $s^{(0)}_{t,k}(m) + \sum_{t'<t,\ k' \in [K]} s^{t',k'}_{t,k}(m)$.

In this paper, we aim to study generative models for (1) under settings such as social media viral streams, where

[1] We sometimes avoid explicitly specifying the sample number m on the counting variables s for ease of reading.

there may exist rare events which invoke disproportionately high endogenous responses. As a representative example, we show a real world plot (Figure 1, left) of cumulative views (normalized) of posts in a social media platform vs time. It is clearly seen that while some of the sequences increase consistently, few others observe sudden jumps in their number of views, which are attributed to such rare events creating huge amounts of response (child events).

3 Existing Approaches

Point processes such and Poisson and Hawkes serve as effective tools for modeling temporal event streams [22, 10], which capture the self-exciting and mutually-exciting behavior among multiple event types. Since we know the causal information for the events in the dataset (1), one may model the counts $s_{t,k}^{t',k'}$ as Poisson random variables: $s_{t,k}^{t',k'} \sim \text{Poisson}\left(\lambda_{t,k}^{t',k'}\right)$, where $\lambda_{t,k}^{t',k'}$ denotes the rate of the arrivals. As simple choices, the rate λ may be modeled as $\lambda_{t,k}^{t',k'} = s_{t',k'} w_{k',k}$ or $\lambda_{t,k}^{t',k'} = s_{t',k'} w_{k',k,t-t'}$, where w denotes the expected offspring rate between the subscripted variables. The latter choice is useful if the dataset (1) satisfies that $s_{t,k}^{t',k'} = 0, \forall t > t' + L$ for any $L > 0$. dataset (1). These models for the rate function fail to explicitly capture the correlation in the arrivals across different event-types, and across different time bins. Another model for the rate function is through combination of multiple basis kernels: $\lambda_{t,k}^{t',k'} = s_{t,k} w_{k',k} \sum_{b \in [B]} \phi_b(t - t')$, where ϕ_b denotes the b^{th} kernel. This modeling choice may also be interpreted as discrete-time equivalent of Hawkes processes [16]. In all these approaches, $\lambda_{t,k}^{t',k'}$ is a deterministic quantity once the models are specified, which limits their expressive power to capture scenarios such as rare influential events, which is the particular setting of interest in this paper.

Clustering Hawkes process: In a continuous-time setting, Du et al. [7] studied Dirichlet Hawkes Process (DHP) in the context of clustering documents in a stream, which was extended by Mavroforakis et al. [19] towards a hierarchical setting (HDHP). In DHP and HDHP, each event is mapped to an existing or new cluster through sampling from a stick-breaking prior, and the future influence of such event is characterized by the properties of the cluster. These modeling choices allow for the instantaneous arrival rate of an event-type k at time $x \in \mathbb{R}_+$, denoted as $\tilde{\lambda}(x, k)$ to be random even after completely specifying the history of all events until the time x. One drawback of these approaches is the continuous-time setting, which limits the scalability. In addition, the clusters do not have any specific structure or interpretation. Whereas, in our case, we wish to model different latent states of the process with a specific structural assumption: the rarer a state is, more is the influence towards future events.

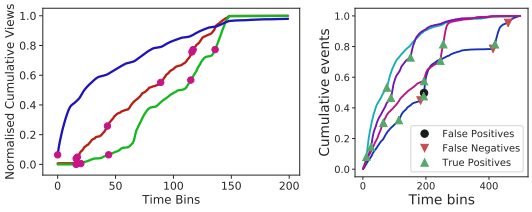

Figure 1: Examples of real world event streams (left) and generated event streams (right) along with predicted rare events. Please refer to Section 6 for more details.

4 BIRTH - A Latent Hawkes Model

In this section, we propose Bayesian dIscRete Time Hawkes process (BIRTH), which addresses the limitations of aforementioned approaches for modeling the dataset (Equation 1). Specifically, we allow the rate $\lambda_{t,k}^{\tilde{t},\tilde{k}}$ to be a random variable, which depends on the assigned latent state of each event.

4.1 Model Formulation

BIRTH follows a discrete time setup similar to that of [16]. We make the following assumptions:

1. Exogenous events $s_{t,k}^{(0)}$ of type k are generated by a Poisson process with constant intensity $\lambda_k^{(0)}\Delta$ with bin size Δ, as $s_{t,k}^{(0)} \sim \text{Poisson}(\lambda_k^{(0)}\Delta)$ (4)
2. Each event $e \in \mathcal{E}_{t,k}$ is associated with a latent state, denoted by $z_e \in [Z]$. We denote $s_{t,k,z} = |\{e \in \mathcal{E}_{t,k} | z_e = z\}|$. Note that $\sum_{z \in [Z]} s_{t,k,z} = s_{t,k}$, the total number of events of type k at time t. $s_{t,k,z}$ is modeled as a Multinomial random variable:
$$[s_{t,k,z}]_{z=1}^{Z} \sim \text{Mult}\left(s_{t,k}, \eta_{t,k}\right), \tag{5}$$
where $\eta_{t,k} \in \mathbb{R}^Z$ denotes the Multinomial class probabilities, which are typically provided as a prior.
3. Each event $e \in \mathcal{E}_{t,k}$, through the mapped latent state z_e, causes future events $e' \in \mathcal{E}_{t',k'}$ for $t' \in [t, t+L], k' \in [K]$, where L denotes the support for endogenous responses. We define
$$s_{t',k'}^{t,k,z} = \left|\{e' \in \mathcal{E}_{t',k'} | e \in \mathcal{E}_{t,k}, (z_e = z), (e \to e')\}\right|.$$
We make the following generative assumption:
$$s_{t',k'}^{t,k,z} \sim \text{Poisson}(\lambda_{t',k'}^{t,k,z}), \tag{6}$$
$$\text{where } \lambda_{t',k'}^{t,k,z} = s_{t,k,z} w_{z,k'} \phi\left(t' - t\right) \Delta.$$

Here, $w_{z,k'}$ is the expected number of offsprings from a latent state z to an event type k', and $\phi(.)$ denotes the time decay kernel, which is assumed to be a probability mass function having a finite support of L time bins.

4. The kernel function $\phi(.)$ is a convex combination of B basis kernels, each of which is a probability mass function with support L given as $\phi(.) = \sum_{b=1}^{B} g_{z,k',b}\phi_b(.)$, where $\sum_b g_{z,k',b} > 0$ and $\sum_b g_{z,k',b} = 1$. Combining this with Equation (6), we see that $\lambda_{t',k'}^{t,k,z} = \sum_{b=1}^{B} \left(s_{t,k,z} w_{z,k'}\ g_{z,k',b}\ \phi_b[t'-t]\Delta\right)$. By the superposition principle of poisson processes [12], this is equivalent to sum of B independent Poisson processes, each corresponding to a basis kernel function ϕ_b. We make this precise by defining the latent variables:

$$s_{t',k'}^{t,k,z,b}(m) = \left|\{e' \in \mathcal{E}_{t',k'}(m) | e \in \mathcal{E}_{t,k}(m), (z_e = z), (e \rightarrow_b e')\}\right| \tag{7}$$

where $e \rightarrow_b e$ means that e causes e' via the Poisson process corresponding to the b^{th} basis kernel ϕ_b. The generative process, in turn, may be written as follows:

$$s_{t',k'}^{t,k,z,b} \sim \text{Poisson}(\lambda_{t',k'}^{t,k,z,b}), \tag{8}$$

$$\lambda_{t',k'}^{t,k,z,b} = s_{t,k,z} w_{z,k'} g_{z,k',b} \phi_b\left(t'-t\right)\Delta. \tag{9}$$

Note that $\sum_{z=1,b=1}^{Z,B} s_{t',k'}^{t,k,z,b} = s_{t',k'}^{t,k}$, which is known.

Model variables: The sample-specific latent variables according to the generative assumptions above are $S_0 = \{s_{t,k,z}\}$, $S_1 = \left\{s_{t',k'}^{t,k,z,b}\right\}$. The global parameters are given by $\Lambda^{(0)} = \{\lambda_k^{(0)}\}$, $W = \{w_{z,k'}\}$, $G = \{g_{z,k',b}\}$ which are provided with the following conjugate priors:

$$\Lambda^{(0)} \sim \text{ Gamma}\left(\alpha^{(\Lambda)}, \beta^{(\Lambda)}\right),$$
$$w_{z,k'} \sim \text{ Gamma}\left(\alpha_z^{(w)}, \beta_z^{(w)}\right), g_{z,k',b} \sim \text{ Dirichlet}\left(\alpha_z^{(g)}\right) \tag{10}$$

where, $\alpha^{(w)}, \beta^{(w)} \in \mathbb{R}^Z$ and $\alpha_z^{(g)} \in \mathbb{R}^B$. $\alpha_z^{(w)} \in \mathbb{R}$ is the z^{th} entry of $\alpha^{(w)}$, similarly for $\beta^{(w)}$. We collectively denote the set of latent variables and global parameters as $\Theta = \left(\Lambda^{(0)}, W, G, S_0, S_1\right)$. The full generative process is described in Algorithm 1.

Likelihood: The likelihood $\mathcal{L}\left(\mathcal{D}(m)\right)$ of the m^{th} data sample (1) according to the generative model is given as follows:

$$\mathcal{L}\left(\mathcal{D}(m)\right) = \prod_{t=1,k=1}^{T,E} \mathcal{L}_{t,k}\left(\mathcal{D}(m)\right), \text{where} \tag{11}$$

$$\mathcal{L}_{t,k}\left(\mathcal{D}(m)\right) = p\left(s_{t,k}^{(0)}(m) | \lambda_k^{(0)}(m)\right) p\left([s_{t,k,z}(m)]_{z=1}^{Z} | s_{t,k}(m)\right) \prod_{t'=t+1,k'=1,z=1,b=1}^{t+L,E,Z,B} p\left(s_{t',k'}^{t,k,z,b}(m) | \lambda_{t',k'}^{t,k,z,b}(m)\right) \prod_{t'=t+1,k'=1}^{t+L,K} \mathbf{1}\left\{\sum_{z,b} s_{t',k'}^{t,k,z,b}(m) = s_{t',k'}^{t,k}(m)\right\},$$

Algorithm 1: Generative Algorithm

Input: Priors for Global model parameters; $\alpha^{(w)}, \beta^{(w)}, \alpha^{\left(\lambda^{(0)}\right)}, \beta^{\left(\lambda^{(0)}\right)}, \alpha_z^{(g)}, \eta_{t,k}$;
Output: $\mathcal{D}$ as defined in Equation 1.
Initialize: Global parameters;
Sample $\lambda_k^{(0)} \sim \text{Gamma}\left(\alpha^{\left(\lambda_k^{(0)}\right)}, \beta^{\left(\lambda_k^{(0)}\right)}\right), \forall k$;
Sample $w_{z,k'} \sim \text{Gamma}\left(\alpha_z^{(w)}, \beta_z^{(w)}\right), \forall z, k'$;
Sample $[g_{z,k',b}]_{b=1}^{B} \sim \text{Dirichlet}\left(\alpha_z^{(g)}\right), \forall z, k'$;
for $t \in [T_m], k \in [K]$ **do**
  Sample $s_{t,k}^{(0)}$ using (4);
  Set $s_{t,k} = s_{t,k}^{(0)} + \sum_{\tilde{t}<t,\tilde{k},\tilde{z},\tilde{b}} s_{t,k}^{\tilde{t},\tilde{k},\tilde{z},\tilde{b}}$;
  Sample $s_{t,k,z}$, $\forall z$, using (5);
  Compute $\lambda_{t',k'}^{t,k,z,b}$, $\forall t'-t \in [1,L], z, b, k'$, as in (9);
  Sample $s_{t',k'}^{t,k,z,b}$ using (8) $\forall t' \in [t+1, t+L], z, b, k'$;
end

and the intensity functions $\lambda_{t'k'}^{t,k,z,b}$ are given by (9). We combine the data likelihood (11) with the priors on the global parameters to arrive at the posterior distribution:

$$p(\Lambda^{(0)}, W, G, S_0, S_1 | \mathcal{D}) = p\left(\Lambda^{(0)}\right) p(W) p(G) \prod_{m=1}^{M} \mathcal{L}\left(\mathcal{D}(m)\right). \tag{12}$$

Discussion:

1. The influence of any event towards future events is assumed to be captured by the assigned latent state. For example, a frequently occurring latent state might be expected to produce smaller number of offsprings than one which is rarer and produces disproportionately high number of offsprings.
2. In (4), the assumption that $\lambda_k^{(0)}$ is time independent is not a limiting factor. One may observe from (11) that the estimation of $\lambda_k^{(0)}$ is completely independent of the remaining parameters and latent variables. In fact, due to the conjugate prior, the posterior distribution of $\lambda_k^{(0)}$ denoted by $p\left(\lambda_k^{(0)} | s_{t,k}^{(0)}\right)$ equals

$$\text{Gamma}\left(\alpha^{\left(\lambda_k^{(0)}\right)} + \sum_{m,t,k} s_{t,k}^{(0)}(m), \beta^{\left(\lambda_k^{(0)}\right)} + \Delta \sum_m T_m\right) \tag{13}$$

The assumptions about $\lambda_k^{(0)}$ may be relaxed without impacting the rest of inference procedures.
3. The assumption in (5) differentiates BIRTH from HDHP [19], where the latter creates clusters among events through a Dirichlet process which can interpreted as topics. BIRTH, instead has an embedded notion through the priors $\left(\alpha_z^{(w)}, \beta_z^{(w)}\right)$ that the latent states capture differences in the offspring produced. Our method differs from HDHP

in the fundamental problem we are tackling, we aim to model the disproportionate responses to different events as latent states.

4. The assumption that ϕ has a finite support of size L is typically valid in many applications. For example, in many popular social media, the feed ranking mechanism which decides the ordering of content on social media gives less priority to older content. Hence, as time elapses the effect of older content decreases, which in other words is equal to lower number of offsprings. The support size L can be chosen based on the statistics from the data such that this assumption holds true.
5. The posterior distribution (12) has non-conjugate factors $p([s_{t,k,z}]_{z=1}^{Z})$ and $p(s_{t'k'}^{t,k,z,b})$. Hence, a full variational inference [25] procedure with mean field distribution for the variables is complicated.

5 Inference

In this section, we derive the inference procedure for BIRTH. As we discussed earlier, estimation of $\Lambda^{(0)}$ may be handled independent of the rest of the variables. The posterior distribution for $\Lambda^{(0)}$ is given in (13). Hence, we shall leave out $\Lambda^{(0)}$ in our remaining discussion.

5.1 Gibbs Sampling

As a simple iterative scheme, Gibbs sampling [2] simulates the global and latent variables of the model by sampling from their corresponding conditional posterior distributions. Recall that $\Theta = \left(\Lambda^{(0)}, W, G, S_0, S_1\right)$ denotes the set of all latent variables and parameters. The conditionals for W, G, S_1 are given are given below:

$$w_{z,k'} | \left(\Theta \setminus w_{z,k'}\right) \sim \text{ Gamma}\left(\alpha_z^{(w)} + \nu_{k'}^{(1)}, \beta_z^{(w)} + \nu^{(2)}\right), \quad (14)$$

$$\text{where } \nu_{k'}^{(1)} = \sum_{m,t,t',k,b} s_{t',k'}^{t,k,z,b}(m), \nu^{(2)} = \sum_{m,t,k} s_{t,k,z}(m)$$

$$g_{z,k'} | \left(\Theta \setminus g_{z,k'}\right) \sim \text{ Dirichlet}\left(\gamma_{z,k'}^{(g)}\right), \quad (15)$$

$$\text{where } \gamma_{z,k'}^{(g)} = \left[\alpha_z^{(g)} + \sum_{m,t,k,t'} s_{t',k'}^{t,k,z,b}(m)\right]_{b=1}^{B}$$

$$\left[s_{t',k'}^{t,k,z,b}\right]_{z,b} | \left(\Theta \setminus \left[s_{t',k'}^{t,k,z,b}\right]_{z,b}\right) \sim \text{Mult}\left(\tilde{a}, \tilde{b}\right), \quad (16)$$

$$\text{where } \tilde{a} = s_{t',k'}^{t,k}, \ \tilde{b} \propto s_{t,k,z} w_{z,k'} \ g_{z,k',b} \ \phi_b(t'-t)$$

For the variables $s_{t,k,z}$, we sample from their conditional posterior after marginalizing out the variables S_1. By Bayes rule, we have the following expression, where the factors $\xi^{(1)}$ and $\xi^{(2)}$ denote the prior for $s_{t,k,z}$ and likelihood of the data $\mathcal{D}$ given $s_{t,k,z}$ respectively:

$$p\left([s_{t,k,z}]_z | W, G, \mathcal{D}, \eta_{t,k}\right) \propto \mathrm{p}\Big([s_{t,k,z}]_z ; s_{t,k}, \eta_{t,k}\Big) \mathrm{p}\Big(\mathcal{D} | s_{t,k,z}, W, G\Big) \quad (17)$$

While the prior on $s_{t,k,z}$ is a Multinomial, the likelihood of the data given $s_{t,k,z}$ is Poisson, which makes the sampling $s_{t,k,z}$ from (17) not straightforward. To address this, we adapt a simple Metropolis Hastings strategy [2] to get the samples, which is detailed in the supplementary material.

5.2 Variational Inference

In this subsection, we propose a hybrid Gibbs-variational inference algorithm similar to that of Mimno et al. [21] to infer the optimal parameters of (12). Variational procedures [3] maintain a distribution about the parameters to be estimated, which is denoted by $q(\Theta; \Psi)$, where Ψ are the parameters of the q-distribution. We restrict q distribution to the families which factor as follows:

$$q(\Theta; \Psi) = \prod_{z,k'} q(w_{z,k'}) \prod_{z,k'} q\left([g_{z,k',b}]_b\right) \prod_{t,k,t',k'} q\left([s_{t,k,z}]_z, \left[s_{t',k'}^{t,k,z,b}\right]_{z,b}\right). \quad (18)$$

Note that (18) is different from the usual mean field assumption, which assumes the q-distribution to be factorized with respect to all variables. Instead, we assume that the latent variables $s_{t,k,z}$ and $s_{t',k'}^{t,k,z,b}$ are coupled with respect to the dimensions z, b in (18). Variational approaches aim to find the best q-distribution (18) which explains the data by solving the following problem:

$$\max_{\Psi} \mathcal{Q}(\Psi) \triangleq \mathbb{E}_q[\log p(\mathcal{D}, \Theta)] - \mathbb{E}_q[\log q(\Theta)]. \quad (19)$$

Here, $\mathcal{Q}$ denotes the evidence lower bound (ELBO), which is a function of the q-distribution.

5.3 Natural Gradient Ascent

For notational convenience, we split Θ into $\Theta_{\mathcal{G}}$ and $\Theta_{\mathcal{V}}$, representing the global parameters and latent variables respectively. Similarly, we denote the corresponding parameters for the q-distribution as $\Psi_{\mathcal{G}}, \Psi_{\mathcal{V}}$. In this notation, $\mathcal{Q}(\Psi_{\mathcal{G}}, \Psi_{\mathcal{V}}) = \mathcal{Q}(\Psi)$. To solve the problem (19), we follow a gradient ascent based approach on $\Psi_{\mathcal{G}}$. Note that (19) can be equivalently rewritten as follows:

$$\max_{\Psi_{\mathcal{G}}} f(\Psi_{\mathcal{G}}), \text{where } f(\Psi_{\mathcal{G}}) = \mathcal{Q}\left(\Psi_{\mathcal{G}}, \Psi_{\mathcal{V}}^*(\Psi_{\mathcal{G}})\right) \triangleq \max_{\Psi_{\mathcal{V}}} \mathcal{Q}(\Psi_{\mathcal{G}}, \Psi_{\mathcal{V}})$$

The gradient of f may be computed as $\nabla f(\Psi_{\mathcal{G}}) = \nabla_{\Psi_{\mathcal{G}}} \mathcal{Q}(\Psi_{\mathcal{G}}, \Psi_{\mathcal{V}}^*(\Psi_{\mathcal{G}}))$, where $\Psi_{\mathcal{V}}^*(\Psi_{\mathcal{G}})$ is the maximiser of $\mathcal{Q}$ (19) with respect to $\Psi_{\mathcal{V}}$ while fixing $\Psi_{\mathcal{G}}$. Instead of the gradient, we use the natural gradient for ascent, which has been shown to have better convergence properties [11].

Computation of the natural gradient: Let $q(w_{z,k'})$ and $q(g_{z,k',b})$ assume a form similar to that of their respective conditional posteriors (14) and (15): $q(w_{z,k'}) = \text{Gamma}(w_{z,k'}; \hat{\alpha}_{z,k'}^{(w)}, \hat{\beta}_{z,k'}^{(w)})$, and $q\left([g_{z,k',b}]_b\right) = \text{Dirichlet}([g_{z,k',b}]_b ; \hat{\alpha}_{z,k'}^{(g)})$. Note that the parameters $[\hat{\alpha}_{z,k'}^{(w)}]_{z,k'}, [\hat{\beta}_{z,k'}^{(w)}]_{z,k'}, [\hat{\alpha}_{z,k'}^{(g)}]_{z,k'}$ comprise $\Psi_{\mathcal{G}}$. We now compute the gradient of $f(\Psi_{\mathcal{G}})$ by expanding (19) only in terms of $\Psi_{\mathcal{G}}$.

Lemma 5.1. *Consider $\mathcal{Q}$ as defined in* (19)*:*

1. *The natural gradient of $\mathcal{Q}$ with respect to $\hat{\alpha}^{(w)}_{z,k'}, \hat{\beta}^{(w)}_{z,k'}$*

$$\begin{bmatrix} \alpha^{(w)}_z + \sum_{t,k,b,t'} \mathbb{E}_q\left[s^{t,k,z,b}_{t',k'}\right] - \hat{\alpha}^{(w)}_{z,k'} \\ \beta^{(w)}_z + \sum_{t,k} \mathbb{E}_q\left[s_{t,k,z}\right] - \hat{\beta}^{(w)}_{z,k'} \end{bmatrix} \tag{20}$$

2. *The natural gradient of $\mathcal{Q}$ with respect to $\hat{\alpha}^{(g)}_{z,k'}$ is given by*

$$\left[\alpha^{(g)}_z + \sum_{t,k,t'} \mathbb{E}_q\left[s^{t,k,z,b}_{t',k'}\right] - \hat{\alpha}^{(g)}_{z,k'}\right]_b \tag{21}$$

Maximizing $\mathcal{Q}$ with respect to Ψ_l: Given $\Psi_{\mathcal{G}}$, it is known that the maximizer of $\mathcal{Q}$ (19) with respect to $\Psi_{\mathcal{V}}$, denoted by $\Psi^*_{\mathcal{V}}$ satisfies the following [3, 11]:

$$\log q(\Theta_{\mathcal{V}}) \propto \mathbb{E}_{q(\Theta_{\mathcal{G}})}\left[\log p\left([s_{t,k,z}]_z, [s^{t,k,z,b}_{t',k'}]_{z,b} | s_{t,k}, s^{t,k}_{t',k'}\right)\right] \tag{22}$$

It is complicated to compute the distribution (22) in closed form. However, we only require $\mathbb{E}_q\left[s_{t,k,z}\right], \mathbb{E}_q\left[s^{t,k,z,b}_{t',k'}\right]$ for the purpose of calculating the gradients using Lemma 5.1 (See (20), (21)). To get these, we may approximate the required quantities by alternatively sampling $s_{t,k,z}$ and $s^{t,k,z,b}_{t',k'}$ from (22). It requires the conditional distributions for $s_{t,k,z}$ and $s^{t,k,z,b}_{t',k'}$, which are given by the following Lemma:

Lemma 5.2. *Consider the distribution* (22)*.*

1. *The conditional distribution for $s^{t,k,z,b}_{t',k'}$ is:*

$$\left[s^{t,k,z,b}_{t',k'}\right]_{z,b} | \left([s_{t,k,z}]_z, \mathcal{D}\right) \sim Mult\left(s^{t,k}_{t',k'}, \left[\delta^{t,k,z,b}_{t',k'}\right]_{z,b}\right), \tag{23}$$

$$where\ \delta^{t,k,z,b}_{t',k'} \propto s_{t,k,z} e^{E_q\left[\log w_{z,k'}\right]}\ e^{E_q\left[\log g_{z,k',b}\right]}\ \phi_b\left[t'-t\right]$$

2. *The conditional distributions for $s_{t,k,z}$ after marginalizing out $s^{t,k,z,b}_{t',k'}$ is given as follows:*

$$q\left([s_{t,k,z}]_z | \mathcal{D}, \eta_{t,k}\right) \propto p\left(s^{t,k}_{t',k'}(m) | s_{t,k,z}(m), E_q\left[W\right], E_q\left[G\right]\right)$$
$$p\left([s_{t,k,z}]_z ; s_{t,k}, \eta_{t,k}\right) \tag{24}$$

Note the similarity of the conditional distributions (23),(24) to that of (16),(17), the only difference being that in the former expectations of the global variables are used, while in the latter their sampled values are used.

Stochastic variational inference From Lemma 5.1, we note that the costliest requirements to compute the gradient are to evaluate $\mathbb{E}_q\left[s_{t,k,z}\right]$ and $\mathbb{E}_q\left[s^{t,k,z,b}_{t',k'}\right]$. This becomes expensive for very large temporal sequences. But it is easy to compute a noisy version of the gradient which leads to a stochastic gradient ascent algorithm as given in Algorithm 2, which requires step sizes ρ_l for each each iteration l. Following Mimno et al. [21], we choose $\rho_l = (\tau_0 + l)^{-\kappa}$, where we choose $\kappa = 0.5$ and $\tau_0 = 10$ in the experiments.

Algorithm 2: Stochastic Variational Inference (SVI)

Data: Binned counts (1), Maximum gradient steps $\mathcal{M}$, Maximum Gibbs Iterations I, Step sizes ρ_t
Result: Optimal variational parameters $\Psi_{\mathcal{G}}$
Initialize variational parameters $\Psi^{(0)}_{\mathcal{G}}$;
for $l \in 1, \ldots, \mathcal{M}$ **do**
  Choose batch $\mathcal{B}_l \subset [T]$;
  for $t \in \mathcal{B}_l$, **do**
    for $i \in [I]$ **do**
      Sample $x^{(i)} = s^{t,k,z,b}_{t',k'}$ using (23);
      Sample $y^{(i)} = s_{t,k,z}$ using (24);
    Compute empirical mean $\hat{s}_{t,k,z}$ using $[x^{(i)}]^I_{i=1}$;
    Compute empirical mean $\hat{s}^{t,k,z,b}_{t',k'}$ using $[x^{(i)}]^I_{i=1}$;
  Approximate the summations
  - $\sum_{t,k,t',k'} \mathbb{E}_q[s^{t,k,z,b}_{t',k'}] \approx \frac{T}{|\mathcal{B}_l|} \sum_{t \in \mathcal{B}_l, k, t', k'} \hat{s}^{t,k,z,b}_{t',k'}$;
  - $\sum_{t,k} \mathbb{E}_q[s_{t',k',z'}] \approx \frac{T}{|\mathcal{B}_l|} \sum_{t \in \mathcal{B}_l, k, z} \hat{s}_{t,k,z}$;

  Compute natural gradient $\nabla_{\Psi_{\mathcal{G}}}(\mathcal{Q})$ (Lemma 5.1);
  Update the variational parameters
  $\Psi^{(l)}_{\mathcal{G}} = \Psi^{(l-1)}_{\mathcal{G}} + \rho_l \nabla_{\Psi_{\mathcal{G}}}(\mathcal{Q})$
Return $\Psi^{(\mathcal{M})}_{\mathcal{G}}$;

5.4 Discussion

We make the following observations comparing Gibbs and SVI (Algorithm 2) procedures to optimize (12).

1. Both the algorithms are similar with respect to sampling the latent variables $s_{t,k,z}$ and $s^{t,k,z,b}_{t',k'}$. In Gibbs iterations, we need to make a complete pass over the whole data before updating the global variables. Hence, SVI (Algorithm 2) scales better for very large datasets.
2. Algorithm 2 is equivalent to an online algorithm when the batches are chosen contiguously from a data stream, thus making it more suitable for large scale online deployments.
3. Gibbs iterations, on the other hand, converges faster with respect to the posterior distribution, and thus is an attractive solution when the complete data can fit into memory.

6 Experiments

BIRTH focuses on modelling the causal relationship which are observed in multidimensional events streams.

This is a less explored problem with significant modeling challenges, as discussed in Sections 1. BIRTH addresses these challenges and provides a unified abstracted framework which can be directly applied in multiple real world setting, for example, in social media content streams, This section evaluates the modelling abilities of BIRTH , by comparing three of our proposed procedures.

1. **Gibbs Inference** (GI): Gibbs Iterations (Section 5.1) which performs multiple passes over the complete data.
2. **Stochastic Variational Inference** (SVI): Iterations as defined in Algorithm 2 which process the data as mini-batches and does multiple passes over the complete data.
3. **SVI-Online**(SVI-O): Processes the complete data once as a stream (sequentially) in mini-batches.
4. **Baseline** (P): A Poisson process baseline as described in Section 3. We model the Poisson rate $\lambda_{t,k}^{t',k'} = s_{t',k'} w_{k',k,l}$. Because the rate is specific to each (k', k, l), this is provides a strong baseline performance. We also compare our method with a variant of Discrete multidimensional Hawkes process which can use causal parent information but its performance is worse as compared to poisson baseline, hence we skip those results in the main paper for better presentation and add them in supplementary Table 1.

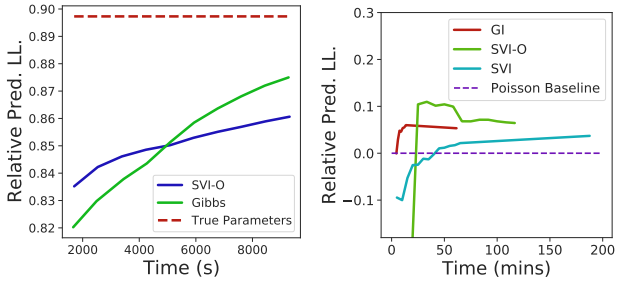

Figure 2: Relative Log Likelihood on validation set. (Left: Synthetic Data, Right: Real Data).

Please note that SVI and GI are comparable in the sense that both of them perform multiple passes over the whole dataset, whereas SVI-O does a single streaming pass. Hence, wherever possible we compare SVI-O with GI to demonstrate its advantage in the online settings.

6.1 Synthetic Experiments

In this section, we compare the inference procedures on a synthetic dataset generated using Algorithm 1. We train BIRTH by generating $M = 100$ event streams with $T = 500$ time bins using the following configurations: $K = 4, Z = 3, B = 10, L = 20$. More details on the data generation are in the supplementary material. In this setting, the dataset fits into the memory and hence both GI and SVI are applicable. We use the learnt model to report the following results.

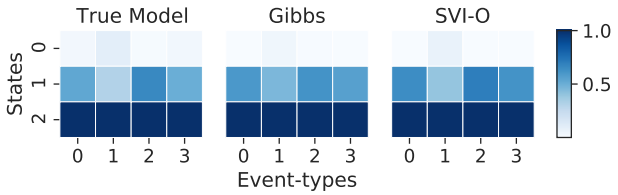

Figure 3: Comparison of True and Recovered w.

Validation Log Likelihood: Using the same configuration as above, we generate a validation dataset of $M = 25$ event streams, in which we evaluate the log-likelihood, given the learnt model parameters. Let $\mathcal{L}_{\mathcal{A}}$ denote the predictive log-likelihood of the data given the parameters learnt using an algorithm $\mathcal{A}$, $\mathcal{A} \in \{\text{GI, SVI-O, P}\}$. We use the relative log-likelihood of the proposed algorithms as a metric for comparison, which is defined as $(\mathcal{L}_{\mathcal{A}} - \mathcal{L}_{\text{P}})/(\#events)$. To ensure fairness of comparison, we compare GI and SVI when their corresponding models were trained for same time duration. The x-axis of Figure 2 refers to the training time, and y-axis refers to the relative log-likelihood, when each of the algorithms are trained for t units. Note that $y = 0$ corresponds to the Poisson baseline and any $y > 0$ means that the model is performing better than the baseline. As expected, both inference procedures have better likelihood of the data than that of the baseline. In addition, Figure 2 also illustrates that, with time, the trained model parameters from both GI and SVI-O approach that of the true model. This attests the fact that SVI-O is a strong procedure in cases where running GI might be infeasible, SVI-O's slower convergence can be attributed to the fact that it is a stochastic algorithm but it comes with the advantage of being an online procedure and memory efficiency: memory requirement of $\mathcal{O}(1)$ as compared to GI which scales linearly with the number of streams and time bins (refer to figure in supplementary).

Structure and True Model Parameter Recovery: Given that the model fits the data well and inference procedure achieves good Likelihood, we try to evaluate the structure of the recovered latent states. Figure 3 plots the expected offspring rate of a latent state ($Z \in \{0, 1, 3\}$) for a particular event type $K \in \{0, 1, 2, 3\}$ for the true model, GI and SVI-O as heatmaps. Here, similar colors denote similar expected offspring rates. We can see that both the procedures recognizes that the second latent state gives rise to the highest number of offsprings, followed by first latent state, whereas the first creates fewer offsprings. The intra-state variations with respect to event types are captured well by both the methods. We go one step

forward and demonstrate an even stronger results for the recovery of the true planted parameters which were used to generate the synthetic dataset. In Figure 4, for the compared algorithms, we plot the normalized error of the recovered model with respect to the true planted parameters versus iterations. As evident, both algorithms achieve better recovery error as iterations progress, while the recovery error of Gibbs falls faster than that of SVI-O.

Locating Rare Event: To illustrate the ability of BIRTH to capture rare influential events, we compare the time bins of the inferred rare events versus the true time bins, which we know in the case of synthetic experiments. For both the algorithms, GI and SVI-O, we obtain the inferred rare events time bins and then plot the True Positives, False Positives (FP) and False Negatives (FN) in Figure 1. It is evident that most of the rare events are predicted correctly but in very few cases we predict a little late due to which we get FN's and FP's in case of SVI-O. This attest the fact that the inference procedures can locate these rare event.

6.2 Real Data Experiments

This sections aims to demonstrate the modelling abilities of BIRTH on a snapshot of a real world social network dataset. The dataset consists of viral event streams, which are the top 0.1% percentile in terms of the number of views accumulated by the stream. The data had events streams with $T = 86K$, $K = 5$, $\Delta = 3600$ seconds. We modelled this dataset by using three latent states ($Z \in \{0, 1, 2\}$). We choose 3 states to depict low, medium, high reproductive rates, which captures all the observations reasonably.

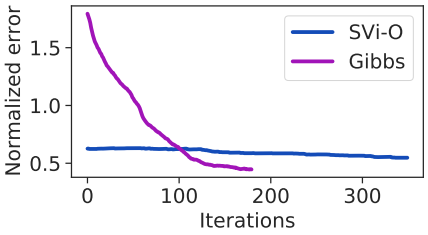

Figure 4: Recovery Error for the global parameters.

Validation Log Likelihood: We compare the Log Likelihood values achieved by GI and SVI and SVI-O inference procedures which is shown in Figure 2. This Figure is to be interpreted in the same way as in Section 6.1. Here, we can see that all three methods achieve better likelihood on a held out validations set as compared to the Poisson baseline.

Locating Rare Events: Similar to section 6.1, we try to locate the rare events (events leading to disproportionate responses) in the stream. We plot the occurrence of events identified as rare (assigned to $Z = 2$) vs the cumulative views attained by the samples (Figure 1). Given that for the real data we don't know the ground truth, hence, we only show inferred rare events. In Figure 1, in *Purple stream*, a rare event occurs in the initial phases of the stream and the video start to accumulate disproportionate number of events after that point. In *Red stream*, the model identifies that few rare events occurred initially and due to that the streams starts to accumulate more events and such events keep on occurring and the stream rises more smoothly, whereas in the *Purple stream* the rate of accumulation of new event decrease as we don't observe another rare event. In *Green Stream*, we observe that rapid surges in the number of events is preceded by rare events. This result further affirms our modelling choice as we are able to provide meaningful explanations for the various growth patterns observed in event stream. Given that the algorithms have some inherent randomness, due to this the latent state occurrence timing might not be always precise, but they can be predicted within reasonable distance. Another point to note is that all of these rare events are not created equal, what this means is that one rare event might create much more offspring as compared to another one. The individual rare events are just tied to each other via the expected offspring count which we demonstrate in Figure 3 but these individual events might have varying level of impact on the streams.

Discussion: Through these set of experiments, we demonstrate that the proposed Generative model and the inference procedures are good choices for modelling and inference on multidimensional event streams with causal parental information. They explain the observed data effectively (i.e., higher Log Likelihood), along with providing structural insights and structure recovery. We also propose a scalable online inference procedure which can be exploited in streaming settings, and show its competitiveness against the other inference procedures proposed. BIRTH allows us to predict occurrence of events which may create disproportionate responses and lead to rapid spread of the stream.

7 Conclusion

The main focus of this paper is to introduce BIRTH , a modelling framework for an abstract setting of multidimensional causal event streams. The paper lays the theoretical foundations for this model, proposes inference routines, and demonstrate the modelling efficacy. BIRTH can potentially be used in various downstream task like prediction if an event stream will go viral or not but we defer such exploration as future works.

References

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

and Liquidity, 1(01):1550005.

[2] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[3] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

[4] Cao, Q., Shen, H., Cen, K., Ouyang, W., and Cheng, X. (2017). Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1149–1158.

[5] Cao, Q., Shen, H., Gao, J., Wei, B., and Cheng, X. (2020). Popularity prediction on social platforms with coupled graph neural networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 70–78.

[6] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936.

[7] Du, N., Farajtabar, M., Ahmed, A., Smola, A. J., and Song, L. (2015). 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.

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

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

[10] Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

[11] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[12] Kingman, J. F. C. (2005). P oisson processes. *Encyclopedia of biostatistics*, 6.

[13] Lemonnier, R., Scaman, K., and Kalogeratos, A. (2016). Multivariate hawkes processes for large-scale inference. *arXiv preprint arXiv:1602.08418*.

[14] Li, C., Ma, J., Guo, X., and Mei, Q. (2017). Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web*, pages 577–586.

[15] Li, H., Ma, X., Wang, F., Liu, J., and Xu, K. (2013). On popularity prediction of videos shared in online social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 169–178.

[16] Linderman, S. W. and Adams, R. P. (2015). Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*.

[17] Liu, S., Yao, S., Liu, D., Shao, H., Zhao, Y., Fu, X., and Abdelzaher, T. (2019). A latent hawkes process model for event clustering and temporal dynamics learning with applications in github. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1275–1285. IEEE.

[18] Martin, T., Hofman, J. M., Sharma, A., Anderson, A., and Watts, D. J. (2016). Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694.

[19] Mavroforakis, C., Valera, I., and Gomez-Rodriguez, M. (2017). Modeling the dynamics of learning activity on the web. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1421–1430.

[20] Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764.

[21] Mimno, D., Hoffman, M. D., and Blei, D. M. (2012). Sparse stochastic inference for latent dirichlet allocation. In *International conference on machine learning*.

[22] Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996). *Stochastic processes*, volume 2. Wiley New York.

[23] Srijith, P., Lukasik, M., Bontcheva, K., and Cohn, T. (2017). Longitudinal modeling of social media with hawkes process based on users and networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 195–202.

[24] Tsur, O. and Rappoport, A. (2012). What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652.

[25] Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

[26] Xu, L., Duan, J. A., and Whinston, A. (2014). Path to purchase: A mutually exciting point process

model for online advertising and conversion. *Management Science*, 60(6):1392–1412.

[27] Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1513–1522.

[28] Zhou, K., Zha, H., and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649.