Hawkes Processes with Stochastic Excitations

Young Lee[*] YOUNG.LEE@NICTA.COM.AU
Kar Wai Lim[†] KARWAI.LIM@ANU.EDU.AU
Cheng Soon Ong[†] CHENGSOON.ONG@ANU.EDU.AU
[*]Data61/National ICT Australia & London School of Economics
[†]Data61/National ICT Australia & Australian National University

Abstract

We propose an extension to Hawkes processes by treating the levels of self-excitation as a stochastic differential equation. Our new point process allows better approximation in application domains where events and intensities accelerate each other with correlated levels of contagion. We generalize a recent algorithm for simulating draws from Hawkes processes whose levels of excitation are stochastic processes, and propose a hybrid Markov chain Monte Carlo approach for model fitting. Our sampling procedure scales linearly with the number of required events and does not require stationarity of the point process. A modular inference procedure consisting of a combination between Gibbs and Metropolis Hastings steps is put forward. We recover expectation maximization as a special case. Our general approach is illustrated for contagion following geometric Brownian motion and exponential Langevin dynamics.

1. Introduction

Motivation. Cascading chain of events usually arise in nature or society: The economy has witnessed that financial meltdowns are often epidemic. For example, the Asian financial crisis swept across Thailand and quickly engulfed South Africa, Eastern Europe and even Brazil. Similarly, criminological research (Bernasco and Nieuwbeerta, 2005) has shown that crime can spread through local environments very rapidly where burglars will constantly attack nearby targets because local susceptibilities are well known to thieves. As another example, in genetic analysis, Reynaud-Bouret and Schbath (2010) looked at the likelihood of occurrences of a particular event along the DNA sequence where 'an event' could be any biological signals occurring along the genomes that tend to cluster together.

The defining characteristic of these examples is that the occurrence of one event often triggers a series of similar events. The Hawkes process, or otherwise known as the self-exciting process, is an extension of Poisson processes that aims to explain excitatory interactions (Hawkes, 1971). What makes the term *self-excitation* worthy of its name is typically not the occurrence of the initial event, but the intensification of further events. We seek to characterize this amplification magnitude, which we call the *contagion parameters*, or *levels of self-excitation*, or simply, *levels of excitation*.

The use of Hawkes processes is not an attempt to describe all features of self-excitation in their correct proportions. Probabilistic modeling inevitably exaggerates some aspects while disregarding others, and an accurate model is one that takes care of the significant aspects and abandons the less important details. Thus, there are often two streams in modeling which are fairly contradictory; on the one hand, the model ought to mimic excitatory relationships in a real world application, and this pulls toward specifying wide families of processes. On the other hand, the model should be manageable and tractable which pulls in the direction of identifying simpler processes in which inference and parameter estimations are feasible.

Adopting the latter view of establishing simpler processes, we present a version of self-exciting processes that permits the levels of excitation to be modulated by a stochastic differential equation (SDE). SDEs are natural tools used to describe rate of change between the excitation levels. Put differently, we attach some indeterminacy to these quantities where we model them as random values over time, rather than being a constant as in the classical Hawkes set-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

ting (Hawkes, 1971; Ozaki, 1979). Our formulation implies that the contagion parameters are random processes thus inheriting tractable covariance structures, in contrast to the set-up initiated by Brémaud and Massoulié (2002) and Dassios and Zhao (2011), where contagion levels are independent and identically distributed (*iid*) random.

Contributions. We present a model that generalizes classical Hawkes and new insights on inference. Our noteworthy contributions are as follows: (1) We propose a fully Bayesian framework to model excitatory relationships where the contagion parameters are stochastic processes satisfying an SDE. This new feature enables the control of the varying contagion levels through periods of excitation. (2) With n denoting the counts of events, we design a sampling procedure that scales with complexity $\mathcal{O}(n)$ compared to a naïve implementation of Ogata's modified thinning algorithm (Ogata, 1981) which needs $\mathcal{O}(n^2)$ steps. (3) A hybrid of MCMC algorithms that provide significant flexibility to do parameter estimation for our self-exciting model is presented. In addition, we describe how to construct two SDEs over periods of unequal lengths and introduce general procedures for inference. (4) We conclude by making explicit deductive connections to two related areas in machine learning; (i) the 'E-step' of the expectation maximization (EM) algorithm for Hawkes processes (Veen and Schoenberg, 2008; Ogata, 1981) and (ii) modeling the volatility clustering phenomenon.

2. Our Model : Stochastic Hawkes

2.1. Review of Poisson and Classical Hawkes Processes

This section recapitulates some pieces of counting process theory needed in what follows. The *Poisson process* is frequently used as a model for counting events occurring one at a time. Formally, the Poisson process with constant intensity λ is a process $N = \{N_t := N(t) : t \geq 0\}$ taking values in $S = \{0, 1, 2, ...\}$ such that: (a) $N_0 = 0$; if $s < t$, then $N_s \leq N_t$, (b) $\mathbb{P}(N_{t+h} = n + m \,|\, N_t = n)$ takes values $\lambda h + o(h)$ if $m = 1$, $o(h)$ if $m > 1$, and $1 - \lambda h + o(h)$ if $m = 0$ where $o(h)$ denotes any function h that satisfies $o(h)/h \to 0$ as $h \to 0$. In addition, if $s < t$, the number $N_t - N_s$ of events in the interval $(s, t]$ is independent to the times of events during $(0, s]$. We speak of N_t as the number of 'arrivals' or 'events' of the process by time t. However, events usually do not arrive in evenly spaced intervals but naturally arrive clustered in time. The *Classical Hawkes process* aims at explaining such phenomenon. It is a point process N whose intensity λ_t depends on the path mirrored by the point process over time. Precisely, the point process is determined by λ_t through the following relations: $\mathbb{P}(N_{t+h} = n + m \,|\, N_t = n)$ takes values $\lambda_t h + o(h)$ if $m = 1$, $o(h)$ if $m > 1$, and $1 - \lambda_t h + o(h)$ if $m = 0$, where $\lambda_t = c_0 + \sum_{i:t>T_i} c_1 \exp(-c_2(t - T_i))$ for positive constants c_0, c_1 and c_2.

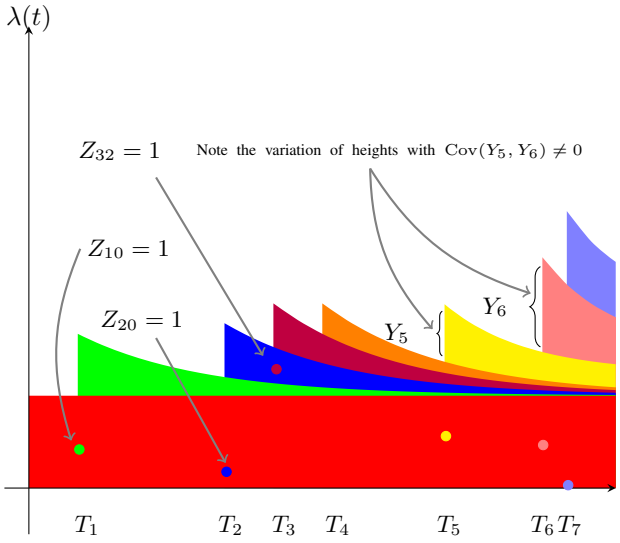


Figure 1. A sample path of the intensity function $\lambda(\cdot)$. First note that the red region represents the base intensity $\hat{\lambda}_0(t)$, which is assumed to be a constant in this diagram. Each colored region, except the red, represents the excitation contributed by each event time. For example, the blue region is contributed by T_2, which in turn is represented by the blue dot (●). Further, note that the blue dot (●) lies in the interior of the red region, indicating that it is an *immigrant*. Mathematically, this is represented by $Z_{20} = 1$. On the other hand, the offspring of the second event T_2 is represented by the maroon dot (●), which is right on top of T_3. We denote this by $Z_{32} = 1$. Observe that this offspring (●) immediately induces another region to be conceived, which is consistently colored in maroon. Stochastic Hawkes is capable of capturing and resembling different levels of contagion and this is evident from the differing heights in Y at the event times T_5 and T_6, where we also allow for a non-zero covariance structure, i.e., $\mathrm{Cov}(Y_5, Y_6) \neq 0$. Finally, it is important to note that the higher the intensity $\lambda(\cdot)$ is, the stronger the rate of decay is. In other words, the gradient is the same for a fixed intensity level, which is a property of the exponential kernel ν.

2.2. Proposed Model Specification and Interpretation

We define our model as a linear self-exciting process $N(t)$ endowed with a non-negative $\mathcal{F}_t$–stochastic intensity function $\lambda(t)$:

$$\lambda(t) = \hat{\lambda}_0(t) + \sum_{i:t>T_i} Y(T_i)\, \nu(t - T_i) \qquad (1)$$

where $\hat{\lambda}_0 : \mathbb{R} \mapsto \mathbb{R}_+$ is a deterministic base intensity, Y is a stochastic process and $\nu : \mathbb{R} \mapsto \mathbb{R}_+$ conveys the positive influence of the past events T_i on the current value of the intensity process. We write $N_t := N(t), \lambda_t := \lambda(t)$ and

$Y_i := Y(T_i)$ to ease notation and $\{\mathcal{F}_t\}$ being the history of the process and contains the list of times of events up to and including t, i.e. $\{T_1, T_2, ..., T_{N_t}\}$. Figure 1 illustrates the different components of our model, which we explain in the following subsections.

2.2.1. Base Intensity, $\hat{\lambda}_0$

This parameter is the base or background intensity describing the arrival of external-originating events in the absence of the influence of any previous events. These events are also known as *exogenous* events. By way of analogy, the base rate is referred to as the 'immigrant intensity' in ecological applications (Law et al., 2009), where it describes the rate with which new organisms are expected to arrive from other territories and colonies. In our case $\hat{\lambda}_0(t)$ is a function of time and takes the form $\hat{\lambda}_0(t) = a + (\lambda_0 - a)e^{-\delta t}$ where $\lambda_0 > 0$ is the initial intensity jump at time $t = 0$, $a > 0$ is the constant parameter, and $\delta > 0$ is the constant rate of exponential decay.

2.2.2. The Contagion Process, $(Y_i)_{i=1,2,...}$

The levels of excitation Y measure the impact of clustering or contagion of the event times. To see this, observe in Equation (1) that whenever Y is high and of positive value, it imposes a greater value to the intensity λ, thus increasing the probability of generating an event in a shorter period of time, thereby causing the clustering phenomena.

We use differential equations to describe the evolution of the levels of excitation. Translating the evolution of contagiousness into the language of mathematics means setting up an equation containing a derivative (or an integral), discussed further below. The changes in the contagion is assumed to satisfy the stochastic differential equation

$$Y_\cdot = \int_0^\cdot \hat{\mu}(t, Y_t)\, dt + \int_0^\cdot \hat{\sigma}(t, Y_t)\, dB_t$$

where B is a standard Brownian motion and $t \in [0, T]$ where $T < \infty$. Different settings of the functionals $\hat{\mu}$ and $\hat{\sigma}$ lead to different versions of SDEs.

An important criterion for selecting appropriate choices of the couple $(\hat{\mu}, \hat{\sigma})$ essentially boils down to how we decide to model the levels of excitation within Stochastic Hawkes. A standing assumption is that the contagion process has to be positive, that is,

Assumption 1 *The contagion parameters* $Y_t > 0, \forall t \geq 0$.

This is necessary as the levels of excitations Y act as a parameter that scales the magnitude of the influence of each past event and subsequently contributes to the quantity λ in Equation (1), which is non-negative.

Some notable examples of the couple $(\hat{\mu}, \hat{\sigma})$ are the Geometric Brownian Motion (GBM): $\hat{\mu} = (\mu + \frac{1}{2}\sigma^2)Y$, $\hat{\sigma} = \sigma Y$ (Kloeden and Platen, 1999; Zammit-Mangion et al., 2012); the Square-Root-Processes: $\hat{\mu} = k(\mu - Y)$, $\hat{\sigma} = \sigma\sqrt{Y}$, (Archambeau et al., 2007; Opper et al., 2010); Langevin equation: $\hat{\mu} = k(\mu - Y)$, $\hat{\sigma} = \sigma$, and their variants (Stimberg et al., 2011; Welling and Teh, 2011; Liptser and Shiryaev, 1978).

Whilst the positivity of Y is guaranteed for GBM, this may not be true for other candidates such as the Langevin dynamics or the Square-Root-Processes. This is because they possess the inherent property that nothing prevents them from going negative and thus may not be suitable choices to model the levels of excitation. Specifically, Square-Root-Processes can be negative if the Feller condition $2k\mu > \sigma^2$ is not satisfied (Feller, 1951; Liptser and Shiryaev, 1978). For real-life applications, this condition may not be respected, thus violating Assumption 1.

To that end, we focus on two specifications of the SDEs, namely the GBM and we tilt the Langevin dynamics by exponentiating it so that the positivity of Y is ensured (Black and Karasinski, 1991):

- Geometric Brownian Motion (GBM):

$$Y_\cdot = \int_0^\cdot \left(\mu + \frac{1}{2}\sigma^2\right) Y_t\, dt + \int_0^\cdot \sigma Y_t\, dB_t \qquad (2)$$

  where $\mu \in \mathbb{R}$ and $\sigma > 0$.

- Exponential Langevin:

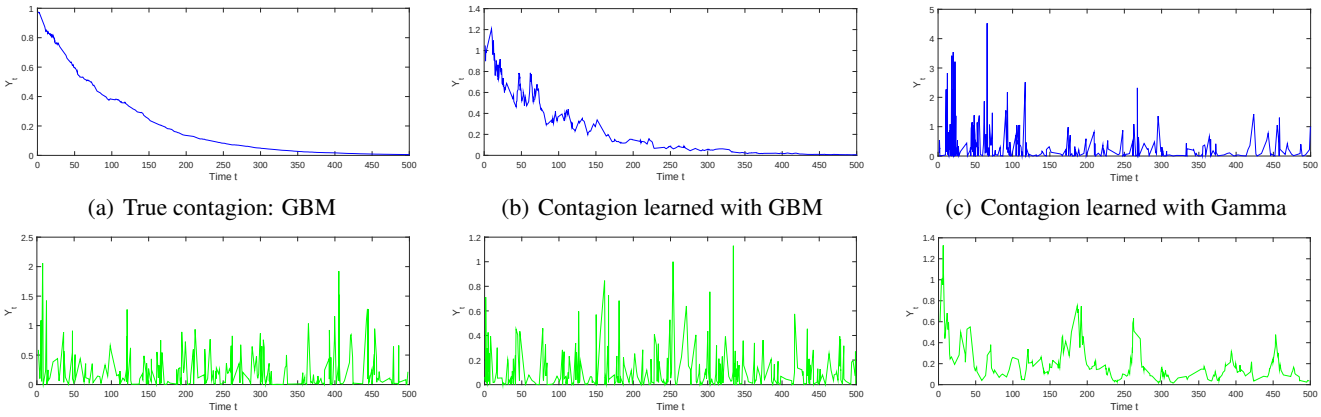
$$Y_\cdot = \exp\left(\int_0^\cdot k(\mu - Y_t)\, dt + \int_0^\cdot \sigma\, dB_t\right) \qquad (3)$$

  where $k, \mu \in \mathbb{R}$ and $\sigma > 0$.

The parameter k for exponential Langevin denotes the decay or growth rate and it signifies how strongly the levels of excitation reacts to being pulled toward the asymptotic mean, μ. For fixed σ and μ, a small value of k implies Y is not oscillating about the mean.

2.2.3. The Sum Product, $\sum_{i:t>T_i} Y(T_i)\, \nu(t - T_i)$.

The product $Y\nu$ describes the impact on the current intensity of a previous event that took place at time T_i. We take ν to be the exponential kernel of the form $\nu(t) = e^{-\delta t}$. Note that δ is being shared between the base intensity $\hat{\lambda}_0$ and the kernel ν to ensure that the process has memoryless properties (see Hawkes and Oakes, 1974; Ozaki, 1979). The memoryless property states the following: given the present, the future is independent of the past. We exploit this property to design our sampling algorithm so that we do not need to track all of its past history, only the present event time matters. This would not have been possible, if we were to use a power kernel (Ogata, 1998), say. In addition, choosing this kernel enables us to derive Gibbs sampling procedures to facilitate efficient inference.



(a) True contagion: GBM (b) Contagion learned with GBM (c) Contagion learned with Gamma

(d) True contagion: Gamma (e) Contagion learned with Gamma (f) Contagion learned with GBM

Figure 2. Versatility of Stochastic Hawkes: Observe that by allowing the level of contagion Y to be a stochastic process satisfying for instance a Geometric Brownian Motion (GBM) makes it possible to reproduce stylized facts of *both* ground truths of GBM and *iid* Gamma variates, see plots (a) & (b) and (d) & (f). On the other hand, it is not possible to perform inference of a more general class of stochastic process, if we were to start with *iid* variables. For example, with the parametrization that Y is *iid* Gamma, one can only reproduce the ground truth when it follows *iid* Gamma but not the stylized facts inherited by a GBM as seen from plots (a) & (c). This can be seen by observing that the path of Y in plot (c) is does not resemble that of the path Y in plot (a).

Summarizing, the intensity of our model becomes:

$$\lambda_t = a + (\lambda_0 - a)e^{-\delta t} + \sum_{i:\, T_i < t}^{N_t} Y_i\, e^{-\delta(t-T_i)}. \quad (4)$$

If we set Y to be a constant, we retrieve the model proposed by Hawkes (1971). In addition, setting $Y = 0$ returns us the inhomogeneous Poisson process. Furthermore, letting $Y = 0$ and $\lambda_0 = a$ simplifies it to the Poisson process.

2.3. The Branching Structure for Stochastic Hawkes

This section presents a generative view of our model that permits a systematic treatment of situations where each of the observed event times can be separated into *immigrants* and *offsprings*, terminologies that we shall define shortly. This is integral to deriving efficient inference algorithms.

We call an event time T_i an *immigrant* if it is generated from the base intensity $a + (\lambda_0 - a)e^{-\delta t}$, otherwise, we say T_i is an *offspring*. This is known as the *branching structure*. It is therefore natural to introduce a variable that describes the specific process to which each event time T_i corresponds to. We do that by introducing the random variables Z_{ij}, where $Z_{i0} = 1$ if event i is an immigrant, and $Z_{ij} = 1$ if event i is an offspring of event j. We illustrate and elucidate the branching structure through Figure 1. For further details on classification of Hawkes processes via the branching structure, refer to Rasmussen (2013) and Daley and Vere-Jones (2003).

2.4. Advantages of Stochastic Contagion

Before we dive into the technical definitions of simulating samples for our point process and performing parameter inference, we illustrate the effect of model choice. The levels of contagion Y can necessarily come in different flavors: be it a constant, as in the standard classical Hawkes, or a sequence of *iid* random variables, or satisfying an SDE.

First we generate levels of contagion Y following a GBM, see Figure 2(a). Performing inference by learning Y as a GBM leads to good estimation as indicated in Figure 2(b). However, if we let Y to be Gamma distributed, we are less able to reproduce the properties of the ground truth, as is evident from Figure 2(c). This is fairly intuitive as the Gamma distribution does not inherit any serial correlation between the samples as they are *iid*, whereas the ground truth does possess correlation structure of a Geometric Brownian Motion.

Proceeding further, this time with the ground truth Y inheriting *iid* Gamma random variables, as illustrated in Figure 2(d). Learning Y as Gamma in our model leads to good inference as illustrated in Figure 2(e). Further, letting Y to be a GBM enables us to learn some properties of the ground truth this time round, see Figure 2(f). Comparing Figures 2(a) & (c) against Figures 2(d) & (f), we conclude that a fairly general formulation for the level of contagion, such as the GBM, is advantageous in recovering stylized facts of an *iid* random levels of contagion, but not vice versa.

2.5. Likelihood Function

This section explicates the *likelihood* with the presence of an SDE and the branching structure within Stochastic Hawkes. The derivation is new and merits discussion here. A key result is that the integrated intensity function $\Lambda_t := \int_0^t \lambda_v \, dv$ can be derived explicitly, that is,

$$\int_0^t \lambda_v \, dv = \int_0^t \left(a + (\lambda_0 - a)e^{-\delta v} + \sum_{i=1}^{N_v} Y_i e^{-\delta(t-v)} \right) dv$$

$$\overset{(i)}{=} \int_0^t a + (\lambda_0 - a)e^{-\delta v} dv + \int_0^t \int_0^v Y_s e^{-\delta(v-s)} dN_s dv$$

$$\overset{(ii)}{=} \int_0^t a + (\lambda_0 - a)e^{-\delta v} dv + \int_0^t \int_s^t Y_s e^{-\delta(v-s)} dv dN_s$$

$$\overset{(iii)}{=} at + \frac{(\lambda - a)(1 - e^{-\delta t})}{\delta} + \frac{1}{\delta} \sum_{i=1}^{N_t} Y_i \Big(1 - e^{-\delta(t - T_i)} \Big)$$

where we note the second term for (i) is a Riemann integral over a stochastic integral with respect to N, (ii) is due to Fubini's Theorem and (iii) follows from the equivalence between stochastic integral and the sum over event times. Consequently, we get the following:

Proposition 1 *Let $\mathcal{T}, \mathcal{Z}, \mathcal{Y}$ be $\{T_i\}$,$\{Z_i\}$ and $\{Y_i\}$ for $i = 1, 2, ..., N_T$ respectively. Assume that no events have occurred before time 0. Further set $\mathcal{J}_i := \{j : 0 < j < i\}$. Then the likelihood function $\mathbb{P}(\mathcal{T} \mid \mathcal{Z}, \mathcal{Y})$ is given by*

$$e^{-\Lambda_T} \prod_{i=1}^{N_T} \left(a + (\lambda_0 - a)e^{-\delta t} \right)^{Z_{i0}} \prod_{j \in \mathcal{J}_i} \left[Y_j e^{-\delta(T_i - T_j)} \right]^{Z_{ij}}.$$

This generalizes the likelihood function of Lewis and Mohler (2011). It can be viewed that this as an alternative version of the likelihood function found in Daley and Vere-Jones (2003) and Rubin (1972) with the presence of the branching structure coupled with a stochastic process Y. The proof of this result can be found in the Supplemental Materials (Section B).

3. Simulation of Stochastic Hawkes

We present a sampling procedure for our Stochastic Hawkes model in Algorithm 1. This algorithm scales with complexity $\mathcal{O}(n)$ compared to a naïve implementation of Ogata's modified thinning algorithm (Ogata, 1981) which requires $\mathcal{O}(n^2)$ steps with n denoting the number of events. Similarly to Ozaki (1979) but differently from Ogata's method, it is noteworthy to mention that our algorithm does not require the stationarity condition for intensity dynamics as long as $T < \infty$.

Algorithm 1 Simulation of Stochastic Hawkes

1. We firstly set $T_0 = 0$, $\lambda_0^{(1)} = \lambda_0 - a$, and given Y_0.
2. For $i = 1, 2, \ldots$ and while $T_i < T$:
   (a) Draw $S_i^{(0)} = -\frac{1}{a} \log U(0, 1)$.
   (b) Draw $u \sim U(0, 1)$. Set $S_i^{(1)} = -\frac{1}{\delta} \log \Big(1 - \delta / \lambda_{T_{i-1}}^{(1)} \log u \Big)$. Note we set $S_k^{(1)} := \infty$ when the log term is undefined.
   (c) Set $T_i = T_{i-1} + \min \left(S_i^{(0)}, S_i^{(1)} \right)$.
   (d) Sample Y_{T_i} (refer to Algorithm 1 in Supplemental Materials)
   (e) Update $\lambda_{T_i}^{(1)} = \lambda_{T_{i-1}}^{(1)} e^{-\delta(T_i - T_{i-1})} + Y_{T_i}$.

The outline of Dassios and Zhao (2013) for simulating Hawkes processes is followed closely and adapted to the present setting. The idea of their algorithm is to decompose the inter-arrival event times into two independent simpler random variables, denoted by $S^{(0)}$ and $S^{(1)}$, with the intention that they can be sampled conveniently. Note however that we also need to sample the levels of self-excitation Y, which is a stochastic process in contrast to *iid* sequences of Y as in Dassios and Zhao (2013).

We seek to find laws that describe the GBM and exponential Langevin dynamics. Applying Itô's formula (see Liptser and Shiryaev, 1978, and Section C in the Supplemental Materials) on $f(y) = \log(y)$ and performing discretization for GBM yields

$$Y_i = Y_{i-1} \exp \left(\mu \Delta_i + \sqrt{\sigma \Delta_i} \, \epsilon_i \right), \tag{5}$$

where Δ_i is introduced as a shorthand for $T_i - T_{i-1}$. Similarly for exponential Langevin, the discretization scheme returns

$$\log Y_i = (\log Y_{i-1})\phi_i + \mu(1 - \phi_i) + \sqrt{\frac{\sigma^2}{2k} \Big(1 - (\phi_i)^2 \Big)} \, \epsilon_i$$

where we define $\phi_i = e^{-k\Delta_i}$, $\epsilon_i \sim N(0, 1)$ is standard normal, and Y_0 is known. Both these expressions now allow us to sample Y_i for all i. We state the following:

Proposition 2 *The simulation algorithm for a sample path of Stochastic Hawkes process is presented in Algorithm 1.*

The proof of this algorithm presented in the Supplemental Materials (Section A).

4. Parameter Inference from Observed Data

We present a hybrid of MCMC algorithms that updates the parameters one at a time, either by direct draws using Gibbs sampling or through the Metropolis–Hastings (MH) algorithm. A hybrid algorithm (Robert and Casella, 2005) combines the features of the Gibbs sampler and the MH algo-

rithm, thereby providing significant flexibility in designing the inference thereof for the parameters within our model.

To see the mechanics of this, consider a two-dimensional parameterization as an illustration. Let θ_A and θ_B be parameters of interest. Assume that the posterior $\mathbb{P}(\theta_B \mid \theta_A)$ is of a known distribution, we can perform inference directly utilizing the Gibbs sampler. On the other hand, suppose $\mathbb{P}(\theta_A \mid \theta_B)$ can only be evaluated but not directly sampled; then, we resort to the use of an MH algorithm to update θ_A given θ_B. The MH step samples from a proposal distribution $\mathbb{Q}\big(\theta'_A \mid \theta_A^{(j)}, \theta_B^{(j)}\big)$ which implies that we draw $\theta_A^{(j+1)} \sim \mathbb{Q}\big(\theta'_A \mid \theta_A^{(j)}, \theta_B^{(j)}\big)$ and that the criteria to accept or reject the proposal candidate is based on the acceptance probability, denoted by $AP\big(\theta_A^{(j+1)}\big)$:

$$\min\left(1, \frac{\mathbb{P}\big(\theta'_A \mid \theta_B^{(j)}\big)\,\mathbb{Q}\big(\theta_A^{(j)} \mid \theta'_A, \theta_B^{(j)}\big)}{\mathbb{P}\big(\theta_A^{(j)} \mid \theta_B^{(j)}\big)\,\mathbb{Q}\big(\theta'_A \mid \theta_A^{(j)}, \theta_B^{(j)}\big)}\right). \tag{6}$$

The hybrid algorithm is as follows: given $\Big(\theta_A^{(0)}, \theta_B^{(0)}\Big)$, for $j = 0, 1, ..., J$ iterations:

1. Sample $\theta_A^{(j+1)} \sim \mathbb{Q}\big(\theta'_A \mid \theta_A^{(j)}, \theta_B^{(j)}\big)$ and *accept* or *reject* $\theta_A^{(j+1)}$ based on Equation (6).
2. Sample $\theta_B^{(j+1)} \sim \mathbb{P}\big(\theta_B \mid \theta_A^{(j+1)}\big)$ with Gibbs sampling.

We proceed by explaining the inference of a simple motivating example in Section 4.1. This is the case when the contagion parameters Y are *iid* random elements. The main inference procedures for Y being stochastic processes can be found in Sections 4.2 and 4.3.

We summarize our MCMC algorithm in Algorithm 2.

4.1. Example: Levels of Excitation Y are *iid* Random

The focus here is on Y to be *iid* random elements with distribution function $G(y), y > 0$. To form a suitable model for the problem under consideration, we propose to model Y as a sequence of *iid* Gamma distribution. This is a slight generalization to the Exponential distribution suggested by Rasmussen (2013) as Gamma distribution contains an additional shape parameter that will help to improve the fitting performance.

The Y_i are assumed to inherit *iid* Gamma distribution with shape τ and scale ω: $\mathbb{P}(y \mid \tau, \omega) \propto y^{\tau} e^{-\omega y}$. We also fix Gamma priors for $\{a, \lambda_0, \delta, \tau, \omega\}$ with hyperparameters $\{(\alpha_m, \beta_m)$ where $m = a, \lambda_0, \delta, \tau, \omega)\}$. Since all branching structure is equally likely *a priori*, we have $\mathbb{P}(\mathcal{Z}) \propto 1$. The posterior for Y_i follows Gamma distribution, $Y_i \mid \cdot \sim \Gamma\big(\tau + \sum_{r=i+1}^{N_T} Z_{ri}\ ,\ \omega + \frac{1-e^{-\delta(T-T_i)}}{\delta}\big)$ which can easily be sampled. Turning to the parameters of Y, we note that Gamma prior on ω gives Gamma posterior, $\omega \mid \cdot \sim \Gamma\big(\alpha_\omega + \tau N_T,\ \beta_\omega + \sum_{i=1}^{N_T} Y_i\big)$

Algorithm 2 MCMC Algorithm For Stochastic Hawkes

1. Initialize the model parameters by sampling from their priors.
2. For all Z_i: Use Gibbs sampler to generate a sequence of Z_i using the posterior distribution defined in Equation (7) with parameters derived in Section 4.2.1.
3. Depending on the choice of SDE so that for all Y_i: sample Y_i using an MH scheme as tabulated in Table 1 in Supplemental Materials (Section D).
4. For a, λ_0 and δ: sample these quantities with an MH scheme as tabulated in Table 1 in Supplemental Materials (Section D).
5. For the contagion parameters μ, σ^2 and k, perform Gibbs sampling using the posterior parameters derived in Section 4.2.2.
6. Repeat steps 2–6 until the model parameters converge or when a fixed number of iterations is reached.

and the acceptance probability for the sampled τ' is given by $\min(1, A(\tau'))$ where $A(\tau')$ takes the form $\big(\omega^{N_T} \prod_{i=1}^{N_T} Y_i\big)^{\tau'-\tau} \big(\frac{\tau'}{\tau}\big)^{\alpha_\tau - 1} \big(\frac{\Gamma(\tau')}{\Gamma(\tau)}\big)^{-N_T} e^{-(\tau'-\tau)\beta_\tau}$.

4.2. Gibbs Sampling

4.2.1. Sampling the Branching Structure $\mathcal{Z}$

The posterior of $\mathcal{Z}$ follows the Multinomial posterior distribution $Z_i \mid \mathcal{T}, \mathcal{Y}, \delta, a, \lambda_0 \sim \text{Multinomial}(\mu_{i\cdot})$ where $\mu_{i\cdot} = \mu_{ij}$ for all j is a probability matrix (each row sum to 1) satisfying

$$\mu_{ij} := \begin{cases} \mathbb{P}(Z_{i0} = 1) = \frac{a + (\lambda_0 - a)e^{-\delta T_i}}{W_i} & \text{if } j = 0 \\ \mathbb{P}(Z_{ij} = 1) = \frac{Y_j e^{-\delta(T_i - T_j)}}{W_i} & \text{if } 0 < j < i \end{cases} \tag{7}$$

where

$$W_i = a + (\lambda_0 - a)e^{-\delta T_i} + \sum_{0<j<i} Y_j\, e^{-\delta(T_i - T_j)} \tag{8}$$

is a normalizing constant. In the Gibbs sampler, we sample new Z_i directly from its posterior.

4.2.2. Sampling the Contagion Parameters

Geometric Brownian Motion. Let $X_i = \log(Y_i / Y_{i-1})$ and $\mathcal{X} = (X_1, X_2, ..., X_{N_T})$. Given that the joint posterior is given by $\mathbb{P}(\mu, \sigma^2 \mid \mathcal{X})$, we take two independent conjugate priors, $\mathbb{P}(\mu) \sim N(\mu_0, \sigma_0^2)$ and $\mathbb{P}(\sigma^2) \sim \Gamma_{\text{Inv}}(\alpha_0, \beta_0)$ where Γ_{Inv} refers to the inverse Gamma distribution. Standard calculations yield the posterior distributions $\mathbb{P}(\mu \mid \sigma^2, \mathcal{X}) \sim N(\mu_*, \sigma_*^2)$ and $\mathbb{P}(\sigma^2 \mid \mu, \mathcal{X}) \sim$

$\Gamma_{\text{Inv}}(\alpha_*, \beta_*)$ where the posterior parameters are given by

$$\mu_* = \frac{\sigma_0^2 \sum_{i=1}^{N_T} X_i + \mu_0 \sigma^2}{\sigma_0^2 \sum_{i=1}^{N_T} \Delta_i + \sigma^2},\ \sigma_*^2 = \left(\frac{\sum_{i=1}^{N_T} \Delta_i}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1}, \tag{9}$$

and

$$\alpha_* = \alpha_0 + \frac{N_T}{2},\ \ \beta_* = \beta_0 + \frac{1}{2} \sum_{i=1}^{N_T} \frac{(X_i - \mu \Delta_i)^2}{\Delta_i}. \tag{10}$$

Exponential Langevin. We take similar priors for μ, σ^2 as in the case for GBM. We further assume that $k \sim N(\mu_k, \sigma_k^2)$. The posterior distributions for μ, σ^2 and k are $N(\hat{\mu}_*, \hat{\sigma}_*^2)$, $\Gamma_{\text{Inv}}(\hat{\alpha}_*, \hat{\beta}_*)$ and $N(\hat{\mu}_k, \hat{\sigma}_k^2)$ with

$$\hat{\mu}_* = \frac{\sigma_0^2 \sum_{i=1}^{N_T} (\log Y_i - \phi_i \log Y_{i-1}) \xi_i + \mu_0 \sigma^2}{\sigma_0^2 \sum_{i=1}^{N_T} \xi_i \phi_i^- + \sigma^2}, \tag{11}$$

as well as

$$\hat{\sigma}_*^2 = \left(\frac{\sum_{k=1}^{N_T} \xi_i \phi_i^-}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1}, \tag{12}$$

$$\hat{\alpha}_* = \alpha_0 + \frac{N_T}{2}, \tag{13}$$

and also

$$\hat{\beta}_* = \beta_0 + \frac{1}{2} \sum_{k=1}^{N_T} \frac{k (\log Y_i - \phi_i \log Y_{i-1} - \mu \phi_i^-)^2}{\phi_i^- \phi_i^+}. \tag{14}$$

Recall that the parameter k expresses the *wildness* of fluctuation about the mean level μ. A small value of k translates to a volatile Y. If we believe that the levels of self-excitations were erratic, which is of particular interest, then we would want a small value for k. This implies that expanding the power series on the exponential function $e^x \approx 1 + x$ where $x := -2k\Delta_i$ to the first order would be sufficient. This is similar in spirit to the Milstein scheme (Kloeden and Platen, 1999) where higher orders of quadratic variations vanish. For an exact sampling of k, one needs to resort to an MH scheme. In most applications, we can even set k to be a constant and do not perform inference for it. Proceeding, we obtain

$$\hat{\mu}_k = \frac{\sigma_k^2 \sum_{i=1}^{N_T} (\log Y_{i-1} - \mu) + \sigma^2 k_0}{\sigma_k^2 \sum_{i=1}^{N_T} \Delta_i (\log Y_i - \mu)^2 + \sigma^2}, \tag{15}$$

$$\hat{\sigma}_k^2 = \left(\frac{\sum_{k=1}^{N_T} \Delta_i (\log Y_i - \mu)}{\sigma^2} + \frac{1}{\sigma_k^2} \right)^{-1}, \tag{16}$$

where we have used the following shorthand $\Delta_i = T_i - T_{i-1}$, $\phi_i = e^{-k\Delta_i}$, $\phi_i^- = 1 - \phi_i$, $\phi_i^+ = 1 + \phi_i$, and $\xi_i = 2k/\phi_i^+$ throughout the calculations.

4.3. Metropolis-Hastings

For the case of Y following the GBM, we propose a symmetric proposal for Y_i with $g(Y_i' \,|\, Y_i) \sim N(Y_i, \sigma_Y^2)$. The posterior of $\mathcal{Y}$ is $\mathbb{P}(\mathcal{Y} \,|\, \mathcal{T}, \mathcal{Z}, \delta, \mu, \sigma^2)$. The acceptance probability AP for Y_i' is $AP(Y_i') = \min(1, A(Y_i'))$ where

$$\begin{aligned} A(Y_i') = \exp \Bigg[& -\frac{1}{\delta}(Y_i' - Y_i)(1 - e^{-\delta(T - T_i)}) \\ & - \frac{1}{2\sigma^2 \Delta_i} \Bigg\{ \left(\log \left(\frac{Y_{i+1}}{Y_i'} \right) - \mu \Delta_i \right)^2 \\ & \quad - \left(\log \left(\frac{Y_{i+1}}{Y_i} \right) - \mu \Delta_i \right)^2 \Bigg\} \mathbb{I}_{\{T_{i+1} < T\}} \\ & - \frac{1}{2\sigma^2 \Delta_i} \Bigg\{ \left(\log \left(\frac{Y_i'}{Y_{i-1}} \right) - \mu \Delta_i \right)^2 \\ & \quad - \left(\log \left(\frac{Y_i}{Y_{i-1}} \right) - \mu \Delta_i \right)^2 \Bigg\} \Bigg] \left(\frac{Y_i'}{Y_i} \right)^{\sum_{r=i+1}^{N_T} Z_{ri} - 1} \end{aligned}$$

where we have defined $T_{N_T+1} = \infty$ and $\sum_r Z_{ri} = 0$ when $i = N_T$.

For the case of a' with symmetry normal proposal, the acceptance probability is $\min\big(1, A(a')\big)$, where

$$\begin{aligned} A(a') = & \left[\prod_{i=1}^{N(T)} \left(\frac{a' + (\lambda_0 - a') e^{-\delta T_i}}{a + (\lambda_0 - a)\, e^{-\delta T_i}} \right)^{Z_{i0}} \right] \left(\frac{a'}{a} \right)^{\alpha_a - 1} \\ & \times \exp \left((a' - a) \left(\frac{1}{\delta} \big(1 - e^{-\delta T}\big) - T - \beta_a \right) \right) \end{aligned}$$

For the inferences of the remaining parameters λ_0, δ, and Y_i for exponential Langevin, the acceptance probabilities are shown in Table 1 in Supplemental Materials.

5. Discussion and Related Work

Reduction to EM. We show that careful selection of specific priors yields posterior probabilities that coincide with the distribution that is taken under the E-step in the EM (expectation–maximization) algorithm methodology launched by Veen and Schoenberg (2008). They utilized the branching structure as a strategy for obtaining the maximum likelihood estimates of a classical Hawkes process which has intensity as in Equation (4) with Y being a constant (ψ). As in their paper, we define the variables u_i associated with the $i-$th event time T_i as $u_i = j$ if event i is caused by event j and $u_i = i$ if the event i is an immigrant event. The unobserved branching structure u_i is treated as the missing data and used to construct an EM algorithm.

The conditional expected value of the complete data log-likelihoood can be written as

$$Q\big(\vartheta;\vartheta^{(q)}\big)=\mathbb{E}\Big[\log(\text{complete data likelihood})|\,\mathcal{F}_T,\vartheta^{(q)}\Big]$$
$$=\mathbb{E}\Bigg[\sum_{i=1}^{N_T}\mathbb{I}_{\{u_i=i\}}\log(a+(\lambda_0-a)e^{-\delta t})-\int_{T_i}^{T}\nu^\dagger(s-T_i)\,ds$$
$$+\sum_{i=1}^{N_T}\sum_{j\neq i}\mathbb{I}_{\{u_i=j\}}\log\nu^\dagger(T_i-T_j)|\,\mathcal{F}_T,\vartheta^{(q)}\Bigg].$$

The following probabilities are used to find an expression for the conditional expected complete data log-likelihood:

$$\mathbb{P}(u_i=j\,|\,\mathcal{F}_{T_i},T_i)$$
$$=\begin{cases}\dfrac{\nu^\dagger(T_i-T_j)}{a^{(q)}+(\lambda_0^{(q)}-a^{(q)})e^{-\delta^{(q)}T_i}+\sum_{j:T_j<T_i}\nu^\dagger(T_i-T_j\,|\,\vartheta^{(q)})} & \text{(a)}\\ \dfrac{a^{(q)}+(\lambda_0^{(q)}-a^{(q)})e^{-\delta^{(q)}T_i}}{a^{(q)}+(\lambda_0^{(q)}-a^{(q)})e^{-\delta^{(q)}T_i}+\sum_{j:T_j<T_i}\nu^\dagger(T_i-T_j\,|\,\vartheta^{(q)})} & \text{(b)}\end{cases}$$

taking value (a) when $0<j<i$ and value (b) when $j=i$.

The kernel used by Veen and Schoenberg (2008) is $\nu^\dagger(t)=\psi e^{-\delta t}$. Observe that this coincides with Equations (7) and (8) in Section 4.2.1 when Y is set to a constant ψ. These probabilities are analogous to the probabilities used to perform the thinning in the modified simulation algorithm (see Ogata, 1981; Farajtabar et al., 2014; Valera and Gomez-Rodriguez, 2015).

The Renewal Equation Governing $\mathbb{E}(\lambda_t)$. We provide an expression for $\mathbb{E}(\lambda_t)$ when Y follows an SDE:

$$\lambda_t \overset{1.}{=} a+(\lambda_0-a)e^{-\delta t}+\int_0^t Y_s e^{-\delta(t-s)}\,dN_s$$
$$\overset{2.}{=} a+(\lambda_0-a)e^{-\delta t}+\int_0^t Y_s e^{-\delta(t-s)}\lambda_s\,ds$$
$$+\int_0^t Y_s e^{-\delta(t-s)}\,d\left(N_s-\int_0^s\lambda_u\,du\right)$$
$$\mathbb{E}[\lambda_t]\overset{3.}{=} a+(\lambda_0-a)e^{-\delta t}+\int_0^t\mathbb{E}[Y_s]\,\mathbb{E}[\lambda_s]e^{-\delta(t-s)}\,ds.$$

where 1. rewrites λ as a stochastic integral, 2. follows from subtracting and adding the mean value process of λ and 3. propagating expectation through the equation.

Other Point Processes. Simma and Jordan (2010) proposed an EM inference algorithm for Hawkes processes and applied to large social network datasets. Inspired by their latent variable set-up, we adapted some of their hidden variable formulation within the marked point process framework into our fully Bayesian inference setting. We have leveraged ideas from previous work on self-exciting processes to consequently treating the levels of excitation as random processes. Linderman and Adams (2014) introduced a multivariate point process combining self (Hawkes) and external (Cox) flavors to study latent networks in the data. These processes have also been proposed and applied in analyzing topic diffusion and user interactions (Rodriguez et al., 2011; Yang and Zha, 2013). Farajtabar et al. (2014) put forth a temporal point process model with one intensity being modulated by the other. Bounds of self exciting processes are also studied in (Hansen et al., 2015). Differently from these, we breathe another dimension into Hawkes processes by modeling the contagion parameters as a stochastic differential equation equipped with general procedures for learning. This allows much more latitude in parameterizing the self-exciting processes as a basic building block before incorporating wider families of processes.

Studies of inference for continuous SDEs have been launched by Archambeau et al. (2007). Later contributions, notably by Ruttor et al. (2013) and Opper et al. (2010), dealt with SDEs with drift modulated by a memoryless Telegraph or Kac process that takes binary values as well as incorporating discontinuities in the finite variation terms. We remark that the inference for diffusion processes via expectation propagation has been pursued by Cseke et al. (2015). We add some new aspects to the existing theory by introducing Bayesian approaches for performing inference on two SDEs, namely the GBM and the exponential Langevin dynamics over periods of unequal lengths.

6. Final Remarks

We extended the Hawkes process by treating the magnitudes of self-excitation as random elements satisfying two versions of SDEs. These formulations allow the modeling of phenomena when the events and their intensities accelerate one another in a correlated fashion.

Which stochastic differential equation should one choose? We presented two SDEs of unequal lengths in this work. The availability of other SDEs in the machine learning literature leaves us the modeling freedom to maneuver and adapt to each real-life application. Each scenario presents quite distinct specifics that require certain amount of impromptu and improvised inventiveness. Finally, a flexible hybrid MCMC algorithm is put forward and connexions to the EM algorithm is spelled out.

7. Acknowledgments

Young Lee wishes to thank Aditya K. Menon for inspiring discussions.

This work was undertaken at NICTA. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice, Bletchley Park, Bletchley, UK, June 12-13, 2006*, pages 1–16.

Bernasco, W. and Nieuwbeerta, P. (2005). How do residential burglars select target areas? A new approach to the analysis of criminal location choice. *British Journal of Criminology*, 45(3):296–315.

Black, F. and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analyst Journal*, pages 52–59.

Brémaud, P. and Massoulié, L. (2002). Power spectra of general shot noises and Hawkes point processes with a random excitation. *Advances in Applied Probability*, 34(1):205–222.

Cseke, B., Schnoerr, D., Opper, M., and Sanguinetti, G. (2015). Expectation propagation for diffusion processes by moment closure approximations. In *arXiv preprint arXiv:1512.06098*.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*. Springer-Verlag New York, 2nd edition.

Dassios, A. and Zhao, H. (2011). A dynamic contagion process. *Advances in Applied Probability*, pages 814–846.

Dassios, A. and Zhao, H. (2013). Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18:1–13.

Farajtabar, M., Du, N., Gomez-Rodriguez, M., Valera, I., Zha, H., and Song, L. (2014). Shaping social activity by incentivizing users. In *NIPS '14: Advances in Neural Information Processing Systems*.

Feller, W. (1951). Two singular diffusion problems. *The Annals of Mathematics*, 54:173–182.

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, pages 89–90.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503.

Kloeden, P. E. and Platen, E. (1999). *Numerical solution of stochastic differential equations*. Applications of Mathematics. Springer, Berlin, New York.

Law, R., Illian, J., Burslem, D. F. R. P., Gratzer, G., Gunatilleke, C. V. S., and Gunatilleke, I. A. U. N. (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, 97(4):616–628.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, pages 1–16.

Linderman, S. W. and Adams, R. P. (2014). Discovering latent network structure in point process data. In *Thirty-First International Conference on Machine Learning (ICML)*.

Liptser, R. S. and Shiryaev, A. N., editors (1978). *Statistics of Random Processes, II*. Springer-Verlag, Berlin-Heidelberg-New York.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

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

Opper, M., Ruttor, A., and Sanguinetti, G. (2010). Approximate inference in continuous time Gaussian-Jump processes. In *Advances in Neural Information Processing Systems 23*, pages 1831–1839.

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

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642.

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.*, 38:2781–2822.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer-Verlag New York.

Rodriguez, M. G., Balduzzi, D., and Schölkopf, B. (2011). Uncovering the temporal dynamics of diffusion networks. In *ICML*, pages 561–568.

Rubin, I. (1972). Regular point processes and their detection. *IEEE Transactions on Information Theory*, 18(5):547–557.

Ruttor, A., Batz, P., and Opper, M. (2013). Approximate Gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems 26*, pages 2040–2048.

Simma, A. and Jordan, M. I. (2010). Modeling events with cascades of Poisson processes. In *UAI*, pages 546–555. AUAI Press.

Stimberg, F., Opper, M., Sanguinetti, G., and Ruttor, A. (2011). Inference in continuous-time change-point models. In *Advances in Neural Information Processing Systems 24*, pages 2717–2725. Curran Associates.

Valera, I. and Gomez-Rodriguez, M. (2015). Modeling adoption and usage of competing products. In *ICDM*, pages 409–418. IEEE.

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

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*.

Yang, S.-H. and Zha, H. (2013). Mixture of mutually exciting processes for viral diffusion. In *ICML*, pages 1–9.

Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., and Sanguinetti, G. (2012). Point process modelling of the Afghan war diary. *Proceedings of the National Academy of Sciences*, 109(31):12414–12419.