Tweedie-Hawkes Processes: Interpreting the Phenomena of Outbreaks

Tianbo Li, Yiping Ke
School of Computer Science and Engineering
Nanyang Technological University, Singapore
tianbo001@e.ntu.edu.sg, ypke@ntu.edu.sg

Abstract

Self-exciting event sequences, in which the occurrence of an event increases the probability of triggering subsequent ones, are common in many disciplines. In this paper, we propose a Bayesian model called Tweedie-Hawkes Processes (THP), which is able to model the outbreaks of events and find out the dominant factors behind. THP leverages on the Tweedie distribution in capturing various excitation effects. A variational EM algorithm is developed for model inference. Some theoretical properties of THP, including the sub-criticality, convergence of the learning algorithm and kernel selection method are discussed. Applications to Epidemiology and information diffusion analysis demonstrate the versatility of our model in various disciplines. Evaluations on real-world datasets show that THP outperforms the rival state-of-the-art baselines in the task of forecasting future events.

Introduction

Self-exciting event sequences are ubiquitous. In such an event sequence, the occurrence of an event will raise the probability of triggering succeeding events. These events could be: posts in social networking sites, cases of an epidemic, comments on a popular movie, or aftershocks following an earthquake. The self-exciting nature often brings about outbreaks of events in a short period of time. People care about why an outbreak happens. Why are certain tweets re-tweeted so many times but not the others? What factors activate an outbreak of a certain epidemic? To answer these questions, an effective and interpretable tool to model and understand outbreaks is needed.

A typical tool for modeling self-/mutually excited data is Hawkes process (Hawkes 1971). Many recent works have explored its application in many disciplines, such as modeling high-frequency transactions (Bowsher 2007), aftershock prediction (Ogata 1988), network inference (Linderman and Adams 2014), recommendation (Du et al. 2015b), etc. Despite its success in many applications, traditional Hawkes processes and most Hawkes-related models are not competent in capturing outbreaks. They mainly suffer from one or many of the following drawbacks:

- *invariant excitation*—the probability of events triggering subsequent ones are either same or i.i.d. distributed;
- *neglecting content/features*—some models only consider the temporal information of the event sequences, but the contents are usually neglected;
- *weak interpretability*—parameters in the model are inaccessible, especially for those combined with neural networks and non-parametric techniques;
- *unrealistic distribution of aggregation*—Taylor's power law (Taylor 1961), which states that the variance of species population density is proportional to a fractional power of the mean, is more natural and common for population and aggregation, whereas some existing works simply adopt Gaussian distribution;
- *failure to sub-criticality*—sub-criticality is a property that the diffusion process produces finite number of events, which many existing works fail to consider.

In this paper, we propose a Bayesian model called *Tweedie-Hawkes Process* (THP). The model parameterizes the excitation parameter in Hawkes process with a Tweedie regression (Jorgensen 1987) over event features, and provides a solution to all the aforementioned drawbacks.

Why Tweedie distribution is more realistic? There are two reasons. First, Tweedie distribution obeys Taylor's power law (Taylor 1961). This law is applicable to many circumstances, such as the spatial distribution of Colorado beetle (Harcourt 1963) and daily turnovers of stocks traded on the NYSE (Fronczak and Fronczak 2010). Hence, equipped with this law, Tweedie distribution is powerful in modeling data that exhibit aggregation (outbreak) phenomena. Second, Tweedie distribution has two important characteristics: *heavy-tail* and *zero-inflation*. Zero-inflation means the probability has a large mass at zero, resulting a natural sparsity. Empirical studies such as (Oestreicher-Singer and Sundararajan 2012) and (Klugman, Panjer, and Willmot 2012), especially in social networks, suggest that many population distributions are heavy-tailed and zero-inflated. In the context of self-excited data, the majority of events are "silent" (i.e., do not have much excitation effect), while a small percentage of events would trigger numerous descendants (in fact, this is how outbreaks are formed). We call the former

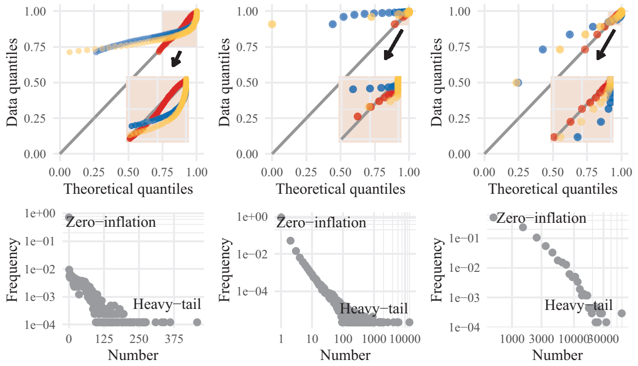


Figure 1: The omnipresence of zero-inflated and heavy-tailed distributions. Columns: Datasets of degrees of the wiki vote network (Leskovec, Huttenlocher, and Kleinberg 2010), the number of retweets in a retweet network (De Domenico et al. 2013) and the amount of claims in vehicle insurance (De Jong, Heller, and others 2008). Top row: Q-Q plots between the actual and thoeretical quantiles of Tweedie (red), Gaussian (blue) and exponential (yellow) distributions. Bottom row: actual distributions of each datasets.

"silent majority" and the latter "vocal minority". Tweedie distribution is able to well describe this phenomena.

Contributions. The contributions of this paper are summarized as follows:

- We propose a Bayesian model combined Hawkes processes and Tweedie distribution called THP, which is able to model outbreaks of self-exciting event sequences and understand the influential factors behind. By leveraging on Tweedie distribution, THP is able to capture the "silent majority" and "vocal minority" in excitation effects, which is also dependent on the features of events.
- We develop an effective mean-field variational EM learning algorithm for model inference. Several theoretical properties of THP, including the sub-criticality and the local optima and convergence of the learning algorithm are also presented. A novel kernel bandwidth selection method is proposed.
- We apply THP to 4 tasks in the experiments, to show the versatility and effectiveness of the model. Two applications to Epidemiology and information diffusion analysis demonstrate the potential of the model. Experimental results also show that THP outperforms the state-of-the-art baselines in data fitting and event prediction.

Preliminaries

Hawkes Process

Hawkes process (Hawkes 1971), a class of point processes, plays a core role in modeling the self- and mutually exciting behavior of events. A Hawkes process $N(t)$ is characterized by its conditional intensity function $\lambda(t|\mathcal{H}_{t\_})$ defined by,

$$\lambda(t|\mathcal{H}_{t\_}) = \lim_{\Delta \to 0} \frac{\mathbb{E}\left[N(t+\Delta) - N(t)|\mathcal{H}_{t\_}\right]}{\Delta},$$

where the history $\mathcal{H}_{t\_}$ is the σ-algebra of events occurring at times up to but not including t. Given a sequence of n events with timestamps $\{t_1, ..., t_n\}$ in an observation window $[0, T]$, the intensity function $\lambda(t|\mathcal{H}_{t\_})$ is given by:

$$\lambda(t|\mathcal{H}_{t\_}) = \mu + \sum_{i:t>t_i} \alpha\phi(t - t_i|h), \tag{1}$$

where $\mu \in \mathbb{R}_+$ is the base parameter controlling the events generated externally, $\alpha \in \mathbb{R}_+$ is the excitation parameter measuring the strength of triggering subsequent events, and $\phi(t|h)$ is the decay kernel function with bandwidth h. In the basic Hawkes, all the events share the same excitation parameter α, regardless of their contents/features.

Tweedie Regression

Tweedie regression is a generalized linear model (GLM) (Nelder and Baker 1972) with the response variable following the Tweedie distribution (Tweedie 1984). Tweedie distribution belongs to the class of the exponential dispersion models (Jorgensen 1987) (EDMs). The probability density function of an EDM is defined by:

$$f(y|\theta, \psi) = c(y|\psi) \exp\left(\frac{y\theta - b(\theta)}{\psi}\right), \quad y \in \mathbb{R}_\psi. \tag{2}$$

Here y is a random variable, θ is called the canonical parameter, and ψ the dispersion parameter. $b(\theta)$ is the cumulant function, and $c(y|\psi)$ is a known function. It is easy to verify that the expectation of y, denoted as η, equals the derivative of $b(\theta)$:

$$\eta \triangleq \mathbb{E}y = b'(\theta). \tag{3}$$

In GLM, the mean of the response variable y is connected to the explanatory variables x in the linear predictor via a smooth and invertible link function g such that $g(\eta) = \boldsymbol{x}'\boldsymbol{\beta}$. Here $\boldsymbol{\beta}$ is the regression coefficients to be inferred. Note that ψ is a nuisance parameter in the estimation of beta. We thus preset ψ and treat it as a constant in this paper.

If y follows a Tweedie distribution, denoted by $y \sim \text{Tweedie}_p(\eta, \psi)$, the variance $\mathbb{V}(y)$ and the mean $\eta = \mathbb{E}(y)$ obeys Taylor's power law (Taylor 1961),

$$\mathbb{V}(y) = \psi\mathbb{E}(y)^p, \tag{4}$$

where $p \notin (0, 1)$.

The Tweedie-Hawkes Process

The idea of the Tweedie-Hawkes process (THP) is to parameterize and randomize the excitation parameter α by the features associated with each event. More specifically, our model defines different α for different events. The intuition is that, different events characterized by different features should have different excitation effects on triggering new events. We achieve this by defining the α in the original Hawkes process as a random variable drawn from a Tweedie distribution. The event features are then naturally incorporated through Tweedie regression. This is essentially to perform a Bayesian treatment on Hawkes. Our THP is detailed as follows.

Consider a sequence of events (i.e., a realization) $\{(t_i, \boldsymbol{x}_i)\}$, $i = 1, \ldots, n$, where $t_i \in [0, T]$ is a timestamp in the observation window and $\boldsymbol{x}_i \in \mathbb{R}^m$ is the corresponding m-dimensional feature vector. Note that we adopt a fixed design setting here, which means that the events associate with the features are served as input. Features are treated as fixed affects here, thus they do not appear in the likelihood function. A discussion on this setting is given in the supplementary material. We denote the timestamp vector by $\boldsymbol{t} = \{t_1, \ldots, t_n\}'$, which is modeled by a Hawkes process with the excitation parameters $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}'$ and the base parameter μ:

$$\boldsymbol{t} \sim \text{Hawkes}(\mu, \boldsymbol{\alpha} | \boldsymbol{\eta}). \tag{5}$$

The log-likelihood of Hawkes process is given by,

$$\ln p(\boldsymbol{t}|\mu, \boldsymbol{\alpha}) = \sum_{i}^{n} \ln \lambda(t|\mathcal{H}_{t_-}) - \int_0^T \lambda(t|\mathcal{H}_{t_-}) dt. \tag{6}$$

Each α_i is drawn from a Tweedie prior distribution, with its mean η_i being a regression over the corresponding feature vector $\boldsymbol{x}_i$,

$$\alpha_i \sim \text{Tweedie}_p(\eta_i, \psi), \tag{7}$$

$$\theta_i = \eta_i^{1-p}/(1-p), \tag{8}$$

$$g(\eta_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{9}$$

Here $g(\eta)$ is a monotonically increasing link function that connects features and the prior distributions. The choice of $g(\eta)$ affects the convergence and sub-criticality of the model. Further details will be discussed later. The prior distribution of α can be written as,

$$\ln p(\boldsymbol{\alpha}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\ln c(\alpha_i; \psi) + \frac{\alpha_i \theta_i - b(\theta_i)}{\psi} \right]. \tag{10}$$

Combining Eq. (6)(10) yields the complete log-likelihood function,

$$\ln p(\boldsymbol{t}, \boldsymbol{\alpha}|\mu, \boldsymbol{\beta}) = \ln p(\boldsymbol{t}|\mu, \boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha}|\boldsymbol{\beta}). \tag{11}$$

Inference

In this section, we develop a variational expectation maximization (VEM) inference algorithm for THP. In the E-step, we approximate the posterior over the hidden variables α_i's using Tweedie variational distributions. In the M-step, we update the model parameters $\boldsymbol{\beta}$ and μ based on the variational distributions. The algorithm is able to achieve local optima and convergence.

E-step. We choose Tweedie distributions as the variational distributions for the latent variables α's. Meanwhile, we factorize the joint distribution of α's by the mean field approximation:

$$q(\tilde{\boldsymbol{\alpha}}) = \prod_{i=1}^{n} q(\tilde{\alpha}_i),$$

where $q(\tilde{\alpha}_i)$ is a Tweedie distribution with corresponding parameter $\tilde{\eta}_i$ (expectation), and the tilde represents variational. Though Tweedie distribution are not the conjugate priors for Hawkes processes, there are two reasons for this choice. First, it results in an evidence lower bound (ELBO) that can be easily approximated by a concave function, guaranteeing local optima. Second, the thorny normalization term $c(y|\psi)$ in the posterior can be canceled out. Fig. 2 gives some results on the true posterior and the approximated one using Tweedie variational distributions. The approximation turns out to be accurate, which justifies our choice of Tweedie as variational distributions.

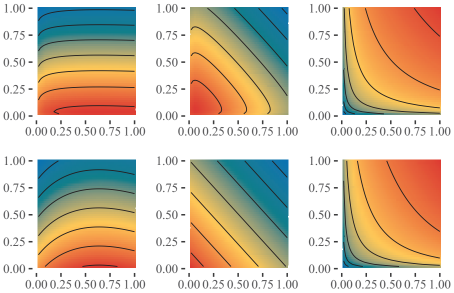


Figure 2: The true posterior over the latent variables (top row) and the approximated one using Tweedie variational distributions (bottom row).

The ELBO for approximating the true posterior distribution of $\tilde{\alpha}_i$ with $q(\tilde{\alpha}_i)$ is given by,

$$\text{ELBO}(\tilde{\alpha}_i) = \int_{\tilde{\alpha}_i} \mathop{\mathbb{E}}_{j \neq i} \ln(\boldsymbol{t}, \boldsymbol{\alpha}|\mu, \boldsymbol{\beta})\, q(\tilde{\alpha}_i)\, \mathrm{d}\tilde{\alpha}_i - \mathop{\mathbb{E}}_{i} \ln q(\tilde{\alpha}_i),$$

where $\mathop{\mathbb{E}}_{j \neq i}$ denotes the expectation over all the other $\tilde{\alpha}_j$'s but $\tilde{\alpha}_i$, leading to a function of the variational variable $\tilde{\alpha}_i$. The full form of the ELBO and detailed derivation is provided in the supplementary material.

This objective function, which is to maximize the ELBO with respect to the variational parameters, however, is not necessarily convex. The second term is a convex-convex fractional function, which spoils the convexity (Benson 2006). We find that the objective function is concave when $p < 2$, ensuring convergence to global optima. We summarize this finding in Lemma 1 and defer the proof to the supplementary material.

Lemma 1 (Concavity)**.** *The ELBO is concave in $\tilde{\eta}_i$ for each i, if $p < 2$.*

M-step. We maximize the expected complete log-likelihood using the variational distribution $q(\tilde{\boldsymbol{\alpha}})$. The Q function with the current parameters μ' and $\boldsymbol{\beta}'$ is given by,

$$Q\left(\mu, \boldsymbol{\beta} | \mu', \boldsymbol{\beta}'\right) = \mathop{\mathbb{E}}_{\tilde{\boldsymbol{\alpha}}} \ln(\boldsymbol{t}, \tilde{\boldsymbol{\alpha}}|\mu, \boldsymbol{\beta}). \tag{12}$$

Applying Eq. (11), the Q function can be decomposed into two parts,

$$Q\left(\mu, \boldsymbol{\beta} | \mu', \boldsymbol{\beta}'\right) = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{\alpha}}} \ln p(\boldsymbol{t}|\mu, \tilde{\boldsymbol{\alpha}})}_{Q(\mu|\mu', \boldsymbol{\beta}')} + \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{\alpha}}} \ln p(\tilde{\boldsymbol{\alpha}}|\boldsymbol{\beta})}_{Q(\boldsymbol{\beta}|\mu', \boldsymbol{\beta}')}, \tag{13}$$

where the first part contains only μ and the second part only $\boldsymbol{\beta}$. It is worth noting that the Q functions here do not contain the current parameters μ' and $\boldsymbol{\beta}'$ outwardly. In fact they are embedded in the variational expectation $\tilde{\eta}_i$. Though the closed-form solutions do not exist, the decoupled Q functions can be easily optimized by various gradient-based methods as they are continuous and smooth. In our implementation, we adopt the Broyden-Fletcher-Goldfarb-Shanno (Avriel 2003) algorithm.

Computational complexity. Given M training sequences of average length N (i.e., N events) with D-dimensional feature vectors, total number of iteration I, the average computation complexity is $\mathcal{O}(I(N^3M + D^2N + D^3))$, which is equivalent to those of the MMEL model (Zhou, Zha, and Song 2013b), GC (Xu, Farajtabar, and Zha 2016) and SLRH (Zhou, Zha, and Song 2013a), in terms of N and I. However, GC involves M basis functions which is much more expensive than ours, and GMHP (Seonwoo, Oh, and Park 2018) employs a strenuous sampling method, which is apparently less efficient. PHP has the same computational complexity as our model, in terms of the number of events and the dimensionality of features. Note that the complexity stated in the PHP paper excludes the feature-related computations and here we include them for fairness.

Theoretical Properties

In this section, we present some theoretical properties of THP. Detailed proofs can be found in the supplementary material.

Sub-Criticality and the Link Function

One desirable property of Hawkes processes is *sub-criticality*, which states that the total progeny of each event is a.s. finite (Vere-Jones 2003). This is an important property as it ensures that the effect of an event will eventually vanish, which is a rule commonly present in many natural phenomena. In this subsection, we show that our THP possesses this important property as long as the link function g satisfies certain conditions. Note that the state-of-the-art model PHP is not sub-critical.

Essentially, the concept of sub-criticality describes a Galton-Watson branching process with a finite total number of events. A Hawkes process can be equivalently interpreted as a collection of branching processes, each centered at an exogenous event (Hawkes and Oakes 1974). Built upon Hawkes, our THP can also be decomposed likewise. More specifically, the events generated by THP come from two sources:

1. $N^{\dagger}$: the process that generates exogenous events;
2. $N^{\ddagger}_{ij}$: the j-th generation of the i-th exogenous event.

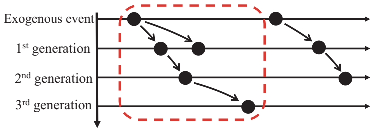


Figure 3: Illustration on the decomposition of THP. Each black dot stands for an event.

An illustration on the decomposition of THP is given in Fig. 3. The event sequence in the red box forms a Galton-Watson branching process. An exogenous event is one triggered by μ whereas endogenous by α. The process $\boldsymbol{N}$ of THP is then the superposition of the above sub-processes:

$$\boldsymbol{N} = \sum_{i=1}^{\boldsymbol{N}^{\dagger}} \sum_{j} \boldsymbol{N}^{\ddagger}_{ij} \tag{14}$$

A Galton-Watson branching process is said to be sub-critical if the expected number of events at each generation is smaller than that at the previous one, as formulated below.

Definition 2 (Sub-Criticality)**.** *A Galton-Watson branching process $N^{\ddagger}_{i}$ is **sub-critical** if $\mathbb{E}\boldsymbol{N}^{\ddagger}_{i,j+1} < \mathbb{E}\boldsymbol{N}^{\ddagger}_{i,j}$ for each generation $j = 1, 2, \cdots$.*

Theorem 3 (Sub-Criticality of THP)**.** *THP consists of a finite number of sub-critical Galton-Watson branching processes if the link function is, (1) invertible, and (2) mapping $(0, 1)$ onto $\mathbb{R}$. Besides, the number of Galton-Watson branching processes $N^{\dagger}$ is a Poisson process of rate μ.*

Convergence Analysis

The learning algorithm presented in the last section is able to achieve local optima and convergence. Theorem 4 states that each iteration of the learning algorithm will consistently increase the likelihood until convergence. The convergence of the model parameters is stated in Theorem 5. ff

Theorem 4 (Local Optima)**.** *For any $k = 1, 2, \cdots$, we have,*

$$\mathcal{L}^{(k+1)} \geq \mathcal{L}^{(k)}, \tag{15}$$

where $\mathcal{L}^{(k)} = \ln p(\boldsymbol{t}|\mu^{(k)}, \boldsymbol{\beta}^{(k)})$ denotes the incomplete log-likelihood of k-th iteration in the learning algorithm of THP.

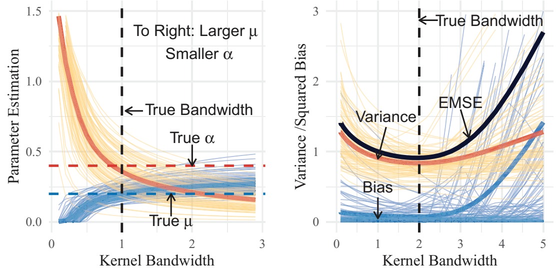
Theorem 5 (Convergence)**.** *If the updating method for $Q(\mu|\mu', \boldsymbol{\beta}')$ and $Q(\boldsymbol{\beta}|\mu', \boldsymbol{\beta}')$ is gradient descent (or Newton-like methods), then as $k \to \infty$, $\|\mu^{(k+1)} - \mu^{(k)}\| \to 0$, $\|\boldsymbol{\eta}^{(k+1)} - \boldsymbol{\eta}^{(k)}\| \to 0$. In particular, the convergence holds for $\boldsymbol{\beta}$ that as $k \to +\infty$, $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\| \to 0$, if either of the following conditions is satisfied: (1) the link function g is uniformly continuous, and (2) the link function g satisfies the sub-critical conditions stated in Theorem 3 and $\eta_i^{(k)} \not\to 0$ or 1 for all i.*

Smoothing Kernel Bandwidth Selection

The selection of kernel bandwidth is an critical problem of Hawkes processes, which is often neglected in existing works. A bad choice of h may result in poor parameter estimations that deviate from true ones. In this part, we are going to propose a solution to the problem.

The proposed method for kernel bandwidth selection is based on the bias-variance trade-off. The main idea is to minimize the expected mean square error (EMSE) on the integrated intensity, which can be decomposed into three parts, the variance, the squared bias, and the irreducible error. More details can be found in the supplementary material.

Fig. 4(a) shows some empirical results on a synthetic dataset. It can be seen that with a true h, the estimation of



(a)

(b)

Figure 4: Illustration on the kernel bandwidth in controlling the trade-offs in exogenous-endogenous events, and in variance-bias. (a): The thin yellow curves show the estimation of α with respect to the kernel bandwidth h, while the thin blue curves represent the estimation of μ. The thick red and thick blue curves represent the respective expectations of the thin curves. (b): Variance (thin yellow over thick red curve) and bias (thin blue curve over thick blue curve) as functions of bandwidth h; a decomposition of the EMSE (thick solid black curve).

α and μ is quite close to the ground truth. Fig. 4(b) shows that the kernel bandwidth governs the trade-off in generating exogenous and endogenous events. A smaller h means a slower decay on the excitation effect, making an event generate more endogenous offspring. This demonstrates the necessity and benefits of selecting the best kernel bandwidth in Hawkes processes.

Related Work

Recent works regarding Hawkes processes make contributions in the following aspects:

Incorporating content features into Hawkes processes. Basic Hawkes process only considers temporal information. Recent works explore to involve the textual information in two fashions: parametric and Bayesian. The first category includes the parametric Hawkes process (PHP) (Li and Zha 2014) model, which parameterizes $\boldsymbol{\alpha}$ with a linear regression on event features. However, the distribution of α in the model is a symmetrical Gaussian and sub-criticality is not necessarily satisfied, as illustrated before. (Tran et al. 2015) proposes a model whose excitations are individuals' participation in communities. The model is similar to PHP, with an application to a clustering task. The other category is more related to our model, where the generation process of features is involved in the model. Typical models also belong to *marked point processes*, such as (Simma and Jordan 2010) and (Seonwoo, Oh, and Park 2018), or mixture models combined Hawkes process and the generation process of features, such as (Yang and Zha 2013), (He et al. 2015) and (Wang et al. 2017). These models always assume that all the events share the same distribution of α (or even same α), which is somehow unrealistic.

Non-parametric Hawkes processes. Another research line is non-parametric Hawkes process. Representative works include the isotonic Hawkes process (Wang et al. 2016), and Hawkes integrated cumulants model (Achab et al. 2017). Neural-based methods (Mei and Eisner 2017), (Li et al. 2018) and (Du et al. 2015b) also gain a lot of attentions recently. These methods, however, suffer from the weakness of interpretability, which do not serve our purpose to find the influencing factors behind event outbreaks.

Bayesian Hawkes processes. As Hawkes processes are a versatile probabilistic model, many recent works apply them to non-parametric Bayesian framework to ease the pain of parameters selection. For example, both DHP (Du et al. 2015a) and DMHP (Xu and Zha 2017) combine Dirichlet process and Hawkes process and apply respective models to clustering tasks. However, they fail to consider the content of the event and assume that the distribution of α only depends on the cluster or nodes (invariant excitation). (He et al. 2015) takes into account the contents and combines Hawkes processes with topic model, but it is a shame that α is treated as a fixed parameter which is only associated to nodes. (Blundell, Beck, and Heller 2012) propose a Bayesian non-parametric model combining Hawkes processes and the infinite relational model. The model claims to discover the implicit social structure by decompose the base intensity term into the products of several factors, but fails to consider the transmission/diffusion of the events between groups, which also follows the basic setting of Hawkes.

Miscellaneous. Another related model is Cox regression (CR) (Cox 1972), which is commonly used in survival analysis for finding the risk factors for a disease. Both CR and our model assume that the intensity function is related to event features. However, CR needs observation on the whole cohort, whereas our model only needs the reported cases. Therefore, our model is suitable when only the positive cases are available. Besides, random graph models (Caron and Fox 2017), share some common components with THP, such as the sparsity of adjacency matrix, heavy-tailed distributions, and many other characteristics in social networks. The main difference is that, random graph-related models are rooted in the generation of adjacency matrix, which is not time-sensitive, whereas our model, with the benefit of Hawkes process, is able to infer the network structure indirectly from the temporal sequences (eg. timelines of Twitter), without knowing any topological structure of the network.

Experiments

In this section, we demonstrate applications and evaluations on both synthetic and real-world datasets. In the first two tasks, we present two applications to Epidemiology and the diffusion of textual information. Then in Task 3, we present that our model has better aggregation of events, which explains why THP is better at capturing outbreaks of events. Last but not least, we test our model on several real-world datasets for forecasting future events in Task 4, which shows our model outperforms the rival baselines.

Task 1: An Application to the Transmission of MERS-CoV

In this task, we apply THP to study the transmission of Middle East respiratory syndrome-coronavirus (MERS-CoV) in

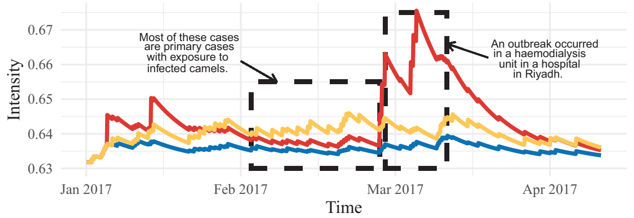


Figure 5: A decomposition of the intensity function with features: `camel milk consumption` (blue), `exposure to MERS-CoV case` (red) and `exposure to camels` (yellow).

Saudi Arabia in 2017. MERS-CoV is a viral respiratory illness that can cause fever, shortness of breath, Pneumonia, and even death. According to the World Health Organization (WHO), approximately 35% of reported MERS patients have died. There are two major routes for transmission: non-human to human transmission, especially from dromedary camels, and human-to-human transmission. However, the transmission patterns of the virus are not fully understood. The aim of this study is to provide some statistical insights for health care workers.

The data is collected from the WHO website, where we study the reported cases in Saudi Arabia in the first 200 days of 2017. The dataset contains 155 cases (i.e., events in our model). We split them into two halves. The first half is for training and the second half for testing. There are three potential risk factors: “exposure to camels”, “camel milk consumption”, and “exposure to MERS-CoV case”. All of them are boolean. They are treated as event features in our model.

We find that “exposure to MERS-CoV case” has the largest regression coefficient, which means that human-to-human transmission is the most statistically important risk factor for MERS-CoV, comparing with the other two factors. Fig. 5 visualizes how the contribution of each factor changes over time by decomposing the intensity function. The respective intensity is calculated by using each feature only. By investigating the decomposed intensities in the figure, we have two findings. First, there are big spikes in intensity from roughly day 60 to day 75. This is supported by an outbreak that occurred in a haemodialysis unit in a hospital in Riyadh between 23 February and 16 March 2017. Second, “exposure to MERS-CoV case” accounts more for this outbreak, as the other two intensity curves are far below the red one. This finding is consistent with the WHO’s observation that several outbreaks occurred primarily due to community transmission within health care settings and households. The cases that were infected by direct or indirect contact with dromedary camels tend to happen individually and occasionally.

Task 2: An Application to Information Diffusion of Textual Contents

Hawkes processes are a potent tool for modeling the dynamics of information and have been applied in many works to the propagation of textual contents, such as (He et al. 2015) and (Seonwoo, Oh, and Park 2018). Due to the omnipresence of the Tweedie distribution, THP greatly enhances Hawkes processes when dealing with text-based cascades in social networks, especially in the following aspects:

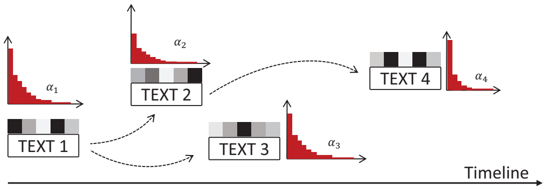


Figure 6: An illustration of the *textual cascade tree* inferred by THP. Each histogram shows the distribution of α for respective text, which is related to the features and the coefficient $\boldsymbol{\beta}$ through Eq. (9). Arrows represent the inferred parent-child relationships, which depend on two factor: (1) temporal distance and (2) textual similarities. Text 1 is a root node. Texts 2, 3 and 4 are descendant of Text 1.

- **Identifying influential texts.** The most important difference between THP and other Hawkes-related models is that in THP, every event has an individual α, which controls the probability of triggering subsequent events. For those events that have larger α, they are more likely to bring about more events. Therefore, η, which is the expectation of α, can be regarded as a indicator of how influential the text is.
- **Popular topic detection.** The Tweedie regression part of THP explains why an event has a larger α (through Eq. (9)). If a bag of words or topic models are used as features, then the value of β represents the contribution of each topic/word, in which
- **Information diffusion modeling based on the latent parent-child relationship of texts.** As a derivative model of Hawkes processes, THP also possesses the branching structure as illustrated in Fig. 3, which provides a method to infer the parent-child relationships among texts. As a consequence, the *textual cascade tree*, which is a directed tree showing the propagation of information in timeline, can be inferred. Fig. 6 demonstrates a toy example.

We test our model on the `MemeTracker` dataset (Leskovec, Backstrom, and Kleinberg 2009). Due to the lack of ground truths (labels), we are not able to really evaluate the model’s performance on finding popular texts or topics. Alternatively, we aim at a prediction task, which is to forecast future dynamics of event sequences. The results are shown in Task 4 with Table 1. Moreover, a case study on the `MemeTracker` dataset is given in the supplementary material.

Task 3: Temporal Aggregation of Events on Synthetic Dataset

Thanks to the Tweedie component, THP is able to generate event data that are more temporally aggregated. We compare with the following baseline models:

- a basic Poisson process (`PP`).

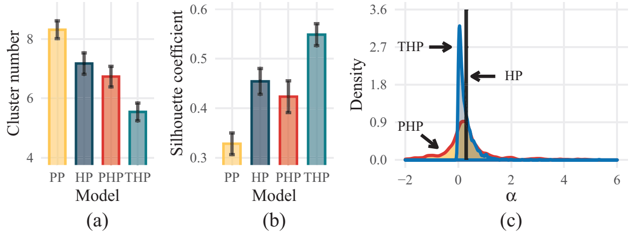


Figure 7: An illustration of the aggregation of events generated by different models. (a)(b): Number of Clusters and Silhouette coefficients of the clustering of events by DBSCAN. It shows that THP has a better aggregation of events in timelines. (c): Comparison of the estimated distributions of α. THP's zero-inflation and heavy-tail can be seen, whereas PHP and HP result a Gaussian distribution and a scalar, respectively.

- a vanilla Hawkes process (HP) (Hawkes 1971);
- Parametric Hawkes process (PHP) (Li and Zha 2014).

We generate 100 datasets using each model (with the same mean of event numbers, and the same mean of α's for HP, PHP, and THP) and use DBSCAN (Ester et al. 1996) to cluster the timestamps in each dataset. As shown in Fig. 7(a) and 7(b), THP obtains the smallest number of clusters and the highest Silhouette coefficient, indicating that it indeed achieves the highest aggregation of events. This is particularly useful for capturing the bursty patterns of real event sequences, where outbreaks of events often occur. Fig. 7(c) further shows the distribution of α learned by different models. HP has a fixed α, while PP has $\alpha = 0$. PHP exhibits a Gaussian-like α distribution. In THP, most α's are around 0, and only a very small percentage of α's have large values. This demonstrates the silent majority and the vocal minority characteristics of THP and explains the highly clustered nature of generated events.

Task 4: Predictions on Real-world Datasets

In this part, we evaluate our model against several state-of-the-art models on 4 real-world datasets in a task of predicting future event sequences. Besides the 3 baselines in Task 3, we also involve 4 more state-of-the-art models to compare with:

- Gaussian Marked Hawkes Processes (GMHP) (Seonwoo, Oh, and Park 2018);
- Sparse Low-rank Hawkes process (SLRH) (Zhou, Zha, and Song 2013a);
- Majorization Minmization Euler-Lagrange algorithm (MMEL) (Zhou, Zha, and Song 2013b);
- Granger Causality for Hawkes (GC) (Xu, Farajtabar, and Zha 2016);

It is worth noting that only PHP, GMHP and our THP incorporate both temporal and feature information, whereas the other several baselines only consider the temporal information. The datasets used for testing are:

- MERS-CoV (MC): the dataset we use in Task 1.
- MemeTracker (MT) (Leskovec, Backstrom, and Kleinberg 2009): the dataset in the Task 2.
- IPTV (Luo et al. 2014): the dataset consists of IPTV viewing events, which records the timestamps and the category that the video belongs to.
- Weeplace (Liu et al.) : This dataset contains the check-in histories of users at different locations. The categories of events include food, education, shops, and 10 others.

The sizes of the datasets we use in the above are: 155, 11275, 2916, 948. We divide each dataset into two parts: 60% as training dataset, and 40% as testing dataset.

Two metrics are used to assess the prediction quality:

- NegLogLik, negative log-likelihood of the test dataset. The likelihood only considers time, with features excluded.
- RMSE, root mean square error of predicting the arrival times of the next N events ($N = 1/5/10$).

A lower NegLogLik and RMSE means the model can better capture the transmission patterns. Due to the page limit, we list the results of log-likelihood in Table 1. For more results, please refer to the supplementary material.

Table 1: The log-likelihood of the predicted future event sequences on various real-world datasets.

Model	MC	MT	IPTV	Weeplace
THP	**-76.344**	**128.455**	**-783.270**	**-1114.906**
PHP	-94.617	126.922	-930.218	-1123.601
GMHP	-80.066	108.048	-969.380	-1187.383
HP	-94.389	102.111	-965.310	-1121.381
GC	-103.287	125.889	-1081.648	-1281.400
SLR	-95.838	52.827	-967.232	-1348.445
MML	-103.693	108.340	-966.795	-1214.991
PP	-104.612	107.982	-1027.998	-1151.463

Conclusion & Discussion

In this paper, we proposed Tweedie-Hawkes Processes, which is powerful in modeling and understanding outbreaks of events. Our model leverages upon the Tweedie distribution in capturing the "silent majority" and "vocal minority" in excitation effects. We showed that the model enjoys a number of theoretical merits and outperforms the state-of-the-art baselines in data fitting and event prediction when tested on several real-world datasets. We also showed the versatility of our model by applying to two tasks in Epidemiology and information diffusion analysis, respectively. It is worth noting that our model is built upon one-dimensional Hawkes processes. Multi-dimensional Hawkes processes can be modified into a THP by setting the dimensionality as a feature. Besides, the distribution of α can be something beyond Tweedie or Gaussian (e.g., binomial distribution). Our model offers a feasible framework to such problem that one can change the Tweedie regression to any other generalized linear models (GLMs).

Acknowledgment

This work is partially supported by the Computational Intelligence Lab (CIL) at the School of Computer Science and Engineering, Nanyang Technological University.

References

Achab, M.; Bacry, E.; Gaïffas, S.; Mastromatteo, I.; and Muzy, J.-F. 2017. Uncovering causality from multivariate hawkes integrated cumulants. *JRML*.

Avriel, M. 2003. *Nonlinear programming: analysis and methods*. Courier Corporation.

Benson, H. P. 2006. Fractional programming with convex quadratic forms and functions. *European Journal of Operational Research*.

Blundell, C.; Beck, J.; and Heller, K. A. 2012. Modelling reciprocating relationships with hawkes processes. In *NIPS'12*.

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

Caron, F., and Fox, E. B. 2017. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(5):1295–1366.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*.

De Domenico, M.; Lima, A.; Mougel, P.; and Musolesi, M. 2013. The anatomy of a scientific rumor. *Scientific reports*.

De Jong, P.; Heller, G. Z.; et al. 2008. Generalized linear models for insurance data. *Cambridge Books*.

Du, N.; Farajtabar, M.; Ahmed, A.; Smola, A. J.; and Song, L. 2015a. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *KDD'15*. ACM.

Du, N.; Wang, Y.; He, N.; Sun, J.; and Song, L. 2015b. Time-sensitive recommendation from recurrent user activities. In *NIPS'15*.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96*.

Fronczak, A., and Fronczak, P. 2010. Origins of taylor's power law for fluctuation scaling in complex systems. *Physical Review E* 81(6):066112.

Harcourt, D. 1963. Population dynamics of leptinotarsa decemlineata (say) in eastern ontario: I. spatial pattern and transformation of field counts. *The Canadian Entomologist*.

Hawkes, A. G., and Oakes, D. 1974. A cluster process representation of a self-exciting process. *Journal of Applied Probability* 11(3):493–503.

Hawkes, A. G. 1971. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)* 438–443.

He, X.; Rekatsinas, T.; Foulds, J.; Getoor, L.; and Liu, Y. 2015. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *ICML'15*.

Jorgensen, B. 1987. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Klugman, S. A.; Panjer, H. H.; and Willmot, G. E. 2012. *Loss models: from data to decisions*.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD'09*. ACM.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010. Predicting positive and negative links in online social networks. In *WWW'10*. ACM.

Li, L., and Zha, H. 2014. Learning parametric models for social infectivity in multi-dimensional hawkes processes. In *AAAI'14*.

Li, S.; Xiao, S.; Zhu, S.; Du, N.; Xie, Y.; and Song, L. 2018. Learning temporal point processes via reinforcement learning. In *NIPS'18*.

Linderman, S., and Adams, R. 2014. Discovering latent network structure in point process data. In *ICML'14*.

Liu, B.; Fu, Y.; Yao, Z.; and Xiong, H. Learning geographical preferences for point-of-interest recommendation. In *KDD'13*. ACM.

Luo, D.; Xu, H.; Zha, H.; Du, J.; Xie, R.; Yang, X.; and Zhang, W. 2014. You are what you watch and when you watch: Inferring household structures from iptv viewing data. *IEEE Transactions on Broadcasting*.

Mei, H., and Eisner, J. M. 2017. The neural hawkes process: A neurally self-modulating multivariate point process. In *NIPS'17*.

Nelder, J. A., and Baker, R. J. 1972. *Generalized linear models*. Wiley Online Library.

Oestreicher-Singer, G., and Sundararajan, A. 2012. Recommendation networks and the long tail of electronic commerce. *Mis quarterly* 65–83.

Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*.

Seonwoo, Y.; Oh, A.; and Park, S. 2018. Hierarchical dirichlet gaussian marked hawkes process for narrative reconstruction in continuous time domain. In *EMNLP'18*.

Simma, A., and Jordan, M. I. 2010. Modeling events with cascades of poisson processes. In *UAI'10*.

Taylor, L. 1961. Aggregation, variance and the mean. *Nature*.

Tran, L.; Farajtabar, M.; Song, L.; and Zha, H. 2015. Netcodec: Community detection from individual activities. In *SDM'15*.

Tweedie, M. 1984. An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, 604.

Vere-Jones, D. 2003. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer.

Wang, Y.; Xie, B.; Du, N.; and Song, L. 2016. Isotonic hawkes processes. In *ICML'18*.

Wang, P.; Fu, Y.; Liu, G.; Hu, W.; and Aggarwal, C. 2017. Human mobility synchronization and trip purpose detection with mixture of hawkes processes. In *KDD'17*.

Xu, H., and Zha, H. 2017. A dirichlet mixture model of hawkes processes for event sequence clustering. In *NIPS'17*.

Xu, H.; Farajtabar, M.; and Zha, H. 2016. Learning granger causality for hawkes processes. In *ICML'16*.

Yang, S.-H., and Zha, H. 2013. Mixture of mutually exciting processes for viral diffusion. In *ICML'13*.

Zhou, K.; Zha, H.; and Song, L. 2013a. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS'13*.

Zhou, K.; Zha, H.; and Song, L. 2013b. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML'13*.