Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Interaction Point Processes via Infinite Branching Model

Peng Lin[§†], Bang Zhang[§], Ting Guo[§], Yang Wang[§], Fang Chen[§]
[§]NICTA, Australian Technology Park, 13 Garden Street, Eveleigh NSW 2015, Australia
[†]School of Computer Science and Engineering, The University of New South Wales, Australia
{peng.lin, bang.zhang, ting.guo, yang.wang, fang.chen}@nicta.com.au

Abstract

Many natural and social phenomena can be modeled by interaction point processes (IPPs) (Diggle et al. 1994), stochastic point processes considering the interaction between points. In this paper, we propose the infinite branching model (IBM), a Bayesian statistical model that can generalize and extend some popular IPPs, *e.g.*, Hawkes process (Hawkes 1971; Hawkes and Oakes 1974). It treats IPP as a mixture of basis point processes with the aid of a distance dependent prior over branching structure that describes the relationship between points. The IBM can estimate point event intensity, interaction mechanism and branching structure simultaneously. A generic Metropolis-within-Gibbs sampling method is also developed for model parameter inference. The experiments on synthetic and real-world data demonstrate the superiority of the IBM.

Introduction

The evolving of our world can be regarded as a series of events, many of which are generally nonindependent. One event may cause or repel the occurrences of others. Examples can be readily found in various of areas. For instance, many biological phenomena compete for local resources, hence demonstrate spatial over-dispersion property. Strong clustering patterns are often observed by seismologists (Marsan and Lengliné 2008) and epidemiologists (Yang and Zha 2013), as earthquakes and epidemics are well known diffusible events. Buy and sell trades in financial markets also arrive in clusters (Hewlett 2006). Information prorogation in social network shows contagious and clustering trait (Yu, Xie, and Sanner 2015). All these events exhibit strong interactive property. Understanding their characteristics can help us categorize, predict and manipulate these events, thereby making positive impacts in our physical and social world.

Despite the high diversity of the aforementioned areas, there are three common tasks for understanding these interactive events: (1) Event intensity estimation, which aims at predicting the number of events for a specific time period. It helps to gain insight into the temporal trends in events. (2) Interaction mechanism estimation, which tries to reveal the triggering or repelling mechanism of events. It provides informative hints for dissemination control and influence manipulation. (3) Branching structure[1] estimation, in which the relationship between events is inferred. It helps to determine the connection of events, understand the underlying causal structure and support event grouping. These three tasks tangle with one another, making the overall problem complex. As a result, most existing approaches only consider one or two of these tasks.

Stochastic point process (Vere-Jones 1988) provides us a generic yet adaptable tool for modeling series of events occurring at random locations and times. It considers a random collection of points falling in some space. When modeling purely temporal events, each point represents the time of an event and the space in which the points fall is simply a portion of the real line. A variety of point processes has been developed with distinct modeling purposes. In this paper, we mainly focus on interaction point processes (IPPs) (Diggle et al. 1994; Ripley 1977) that model not only the generation of points but also their interactions. Specifically, a Bayesian statistical model, which can generalize and extend some popular IPPs, *e.g.*, Hawkes process (Hawkes 1971; Hawkes and Oakes 1974), is proposed with the consideration of the aforementioned three tasks.

Many statistical methods exist for modeling events in spatial and temporal space, and most of them make an exchangeability assumption on a certain component of the overall model (Orbanz and Roy 2013). For instance, random walk models (Bacallado et al. 2013) and the infinite hidden Markov model (Beal, Ghahramani, and Rasmussen 2001), which are widely used for discrete time series, assume Markov exchangeability (Diaconis and Freedman 1980; Zabell 1995) implying that the joint probability only depends on the initial state and the number of transitions. Lévy process, the continuous time analog of random walk and the foundation of many other widely used continuous time models, assumes exchangeability over increments. Another popular example is dependent Dirichlet process (DDP) (MacEachern 1999) which is marginally exchangeable (Orbanz and Roy 2013). It means that DDP is a random measurable mapping whose output is an exchange-

[1] The formal definition of branching structure will be given later. It can be understood as relationship between events for now.

able random structure.

However, in general, the observed events in spatial and temporal space are seldom exchangeable, especially for their dependencies. Distance dependent Chinese Restaurant process (ddCRP) (Blei and Frazier 2011) is a simple yet flexible class of distributions over partitions allowing non-exchangeability. It can be used for directly modeling dependencies between data points in infinite clustering models and the dependencies can be across space and time. In this paper, we adapt the ddCRP as a prior over the branching structure of spatial and temporal events. With its support, a Bayesian statistical model is proposed treating IPP as a mixture of basis point processes (bPPs). It allows discovery of a potentially unbounded number of mixing bPPs, while simultaneously estimating branching structure. We therefore call our approach the infinite branching model (IBM).

The IBM is also related to the infinite relational model (IRM). The IRM aims at inferring meaningful latent structure within observed graph or network. An unbounded number of blocks of nodes with similar behavior can be automatically revealed with the support of the CRP prior on node partitions. But the IBM is interested in discovering the implicit branching structure of a collection of spatial and temporal points based on their positions and distances in space. In terms of the adopted prior for partition, the IBM can be regarded as a distance dependent version of IRM, but for discovering latent branching structure in spatial and temporal space.

Interaction Point Processes

Interaction point process[2] (Diggle et al. 1994; Ripley 1977) is a broad range of stochastic point processes that can model various of interaction mechanisms, *e.g.*, Hawkes process (Hawkes 1971; Hawkes and Oakes 1974), cascades of Poisson processes (Simma and Jordan 2010) and Neyman-Scott process (Neyman and Scott 1958). The work in (Adams 2009) provides a brief summary of some popular IPPs. In this section, we use Hawkes process as an illustration to introduce IPP and show that many popular IPPs can be defined via a Poisson cluster process (Hawkes and Oakes 1974) inspired by which we can later define a mixture model with Poisson processes as bPPs for generalizing and extending these IPPs.

Hawkes Processes

Hawkes process is one of the most general and flexible IPPs. Its formal definition can be given via conditional intensity function. Let $X = \{t_i\}_{i=1}^N$ be a stochastic point process on temporal space, where $t_i \in R$ indicates the time of point. Hawkes process is a family of point processes having the following form of conditional intensity function:

$$\lambda(t) = \mu(t) + \sum_{t_i < t} \alpha\beta(t - t_i). \tag{1}$$

Function $\mu(t)$ is a non-negative function on R, representing immigrant intensity. Variable α is a non-negative parameter representing total offspring intensity. Function $\beta(t)$ is a density function defined on $[0, \infty)$, indicating normalized offspring intensity. Typical normalized offspring intensities are in decay function form, *e.g.*, exponential decay function and logistic decay function. Thus, we can see that the triggering effect of a point appears immediately after its occurrence and quickly decays in certain ways, thereby showing clustering patterns. The product $\alpha\beta(t)$ represents the offspring intensity. The meaning of the intensity and parameter names will become clear in the following. It is worth noting that we use $\lambda(t)$ to represent intensity function conditioned on previous points with the consideration of notation simplicity.

As discussed in (Hawkes and Oakes 1974), Hawkes process can also be viewed equally as a Poisson cluster process that is constituted by a collection of Poisson processes following a certain branching structure. There are two types of points in a Poisson cluster process, immigrant and offspring. The generative procedure of points for a Poisson cluster process can be described as following: (1) The immigrant points $t_i \in I$ are generated via a Poisson process with an immigrant intensity $\mu(t)$. (2) Every immigrant point t_i can generate a cluster of offspring points and the clusters are independent. (3) Within each cluster, points are organized in generations. Generation 0 is simply the immigrant point itself. Every point t_j of a generation can recursively generate a Poisson process O_j with an offspring intensity $\alpha\beta(t - t_j)$, forming the next generation. (4) Finally, Poisson cluster process is the combination of all points.

If a point t_j is generated by a Poisson process O_i, namely $t_j \in O_i$, then we say that point t_j is a child of point t_i and point t_i is the parent of point t_j. The collection of all the parent-child relationships forms the branching structure, denoted by $C = \{c_j\}$. $c_j = i$ means point j is the child of point i and $c_j = j$ means point j is an immigrant point. It is worth noting that for traditional Hawkes process, the offspring intensities are the same for all the points. But we can extend it by allowing different offspring intensities for different clusters. The details will be described in the proposed method. Besides, cluster Poisson process is recursively defined, which means more than one generation of descendants can be generated.

Other Types of Interaction Point Processes

While Hawkes processes address an important case in which an occurrence of a point can cause additional points in near future, there exist other types of IPPs. Interestingly, we can find that, within the Poisson cluster process framework, many IPPs with distinct interaction mechanisms can be defined by choosing different offspring intensity functions. Due to the space limit, we only make a few examples to illustrate it. For instance, cascades of Poisson processes (Simma and Jordan 2010) can be defined as a Poisson cluster process in which offspring intensity function considers all the previous points in its previous generation instead of just its parent. Neyman-Scott process (Neyman and Scott 1958) is a Poisson cluster process that only allows one generation of offspring. For repulsive point processes (Adams 2009; Snoek, Zemel, and Adams 2013), which show inhabitation behaviors, intensity increment is suppressed once a point

[2] Specifically in this work, we only consider pair-wise interaction point processes.

occurs and released in certain ways when the point is far away. Thus, they can be defined as Poisson cluster processes via, for instance, a Gaussian or Weibull shape offspring intensity. As discussed in (Simma and Jordan 2010), periodic activity can also be modeled by Poisson cluster process by using a step function of time as the offspring intensity. Moreover, it is possible to obtain complex interaction mechanisms by defining offspring intensity as a mixture of base intensities as described in (Hawkes 1971; Zhou, Zha, and Song 2013). Hence, all these IPPs can be unified and generalized with the support of the Poisson cluster process.

The Infinite Branching Model

Inspired by the Poisson cluster process, in this work, we propose a Bayesian statistical model, the IBM, that generalizes IPPs as a mixture of Poisson processes. The key component of the IBM is a distance dependent prior over branching structure of points. As mentioned in the introduction section, most of the statistical models designed for spatiotemporal events assume exchangeability, which is unrealistic for modeling point dependencies. Hence, we adapt the ddCRP, a class of non-exchangeable distributions over partitions, as a prior of branching structure.

The ddCRP is a generalization of the Chinese restaurant process (CRP) that is an exchangeable prior on partitions for many popular Bayesian nonparametric models. Unlike CRP, the ddCRP assumes non-exchangeability of data. The order of data affects the distribution of partition structures. It provides a data clustering scheme via the following metaphor. Suppose there is a Chinese restaurant that has an infinite number of tables. A sequence of customers enter and select a table to sit. Customer chooses a table to sit via customer assignment. A customer is either assigned to another customer with probability proportional to a decay function output depending on their distance, or assigned to herself or himself with probability proportional to a concentration parameter. Customer always sits with her or his assigned customer. Thus, table assignment can be obtained via customer assignment as a byproduct.

The direct modeling of customer relationships and its non-exchangeability property make the ddCRP suitable as a prior for branching structure of stochastic points in spatiotemporal space. In the IBM, customers represent points, tables represent point clusters. Customer assignments represent point connections. We say point j is the child of point i (or point j is triggered by point i) if point j is assigned with point i. The distribution of point assignment can be formally described as:

$$p(c_j = i|\eta, f, D) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \eta & \text{if } i = j, \end{cases} \qquad (2)$$

where c_j indicates the point assignment for point j, d_{ij} is the distance between the points i and j, D is the matrix defining pair-wise point distances, and $f(\cdot)$ is a function that mediates how the distance affects the probability of point connection, *e.g.*, window decay function. It reflexes the prior belief of branching structure. It also makes sure that a point can only be assigned to a previously occurred point. A point is an immigrant if it is assigned to itself, and it is an offspring otherwise. Hence, the concentration parameter η controls how likely a point is an immigrant. As we can see that the point type can be determined by point assignment. Point clustering can also be obtained via point assignment indirectly. As in the CRP, each point cluster is endowed with a specific point generation scheme. It is also worth noting that the overall collection of point assignments $C = \{c_j\}$ can now equally represent the branching structure C as described in the previous section. Thus, the ddCRP can be used as a distribution over branching structures. In the following, we use $R(c_{1:N})$ to represent the mapping from point assignment to point cluster assignment, $R^*(c_{1:N})$ to represent the immigrant of the corresponding cluster, and $R^{'}(c_{1:N})$ to represent the offspring of the corresponding cluster.

With the support of the branching structure prior, the IBM can be formally defined. Unlike traditional Hawkes process in which all point clusters share the same offspring intensity, the IBM can allow different offspring intensities for different clusters, which grants more flexibility for modeling real-world events. For defining a concrete model, we assume both immigrant intensity and total offspring intensity are constant variables drawn from exponential distributions. Normalized offspring intensity is in exponential distribution form, $\beta(t) = \lambda^\beta \exp(-\lambda^\beta t)$, with λ^β as its inverse scale parameter drawn from a Gamma distribution. The IBM can be described as following for generating a sequence of points $\{t_i\}$:

1. Sample immigrant intensity $\mu \sim \text{Exponential}(\lambda^\mu)$.
2. Sample t_1 from $\mathcal{PP}(\mu)$, sample its total offspring intensity $\alpha_1 \sim \text{Exponential}(\lambda^\alpha)$ and sample inverse scale parameter for its normalized offspring intensity $\lambda_1^\beta \sim \text{Gamma}(\alpha^{'}, \beta^{'})$.
3. For $n > 1$:
   (a) Sample $t_n > t_{n-1}$ from $\mathcal{PP}(\mu + \sum_{i=1}^{n-1} \alpha_i \beta_i (t - t_i))$
   (b) Sample point assignment $c_n \sim ddCRP(\eta, f, D)$. It indirectly determines cluster assignment and point types: $R(c_n)$, $R^*(c_n)$ and $R^{'}(c_n)$.
   (c) If t_n is an offspring, then set $\alpha_n = \alpha_{R(c_n)}$ and $\lambda^\beta = \lambda^\beta_{R(c_n)}$. Otherwise, for a new cluster, sample its total offspring intensity $\alpha_{R(c_n)} \sim \text{Exponential}(\lambda^\alpha)$ and sample inverse scale parameter for its normalized offspring intensity $\lambda^\beta_{R(c_n)} \sim \text{Gamma}(\alpha^{'}, \beta^{'})$.

In the above, λ^μ, λ^α, $\alpha^{'}$ and $\beta^{'}$ are hyper-parameters. $\mathcal{PP}(\cdot)$ indicates a Poisson process. Samples can be drawn from an inhomogeneous Poisson process by utilizing a thinning process, a point process variant of rejection sampling. Specifically, the Ogata's modified thinning (Ogata 1981) can be used, as summarized by the Algorithm 7.5.IV in (Vere-Jones 1988). The model can be readily simplified for mimicking the traditional Hawkes process. Although we adopt exponential distribution form for normalized offspring intensity, other distribution forms or combinations can be used for modeling different interaction mechanisms, *e.g.*, spatiotemporal interaction. It has been noted that the CRP can be regarded as a special case of the ddCRP. As a result, the branching structure prior in the IBM can become exchangeable in terms of clustering when the ddCRP prior degrades to the CRP. It means that the probability of a point belonging to a cluster only depends on the number of points that are already in the cluster.

Hierarchical Model

It is always desirable to discover latent hierarchical structure from data. For IPPs, it is beneficial to reveal the relationship between point clusters. For instance, finding similar clusters of buy and sell trades in financial market can be insightful for making trading strategy. Hence, we extend the IBM to a hierarchical model in which similar point clusters can form a hyper-cluster sharing the same offspring intensity. For defining a concrete extension, we again assume μ and α are constant variables drawn from exponential distributions. But, for this time, we let normalized offspring intensity be in Weibull distribution form for showing its capability of capturing different interaction mechanism: $\beta(t) = \left(k^{\beta}/\lambda^{\beta}\right)\left(t/\lambda^{\beta}\right)^{k^{\beta}-1} e^{-(t/\lambda^{\beta})^{k^{\beta}}}$. The hierarchical model is described as follow:

1. Sample immigrant intensity $\mu \sim$ Exponential(λ^{μ}).
2. Sample t_1 from $\mathcal{PP}(\mu)$, sample its total offspring intensity $\alpha_1 \sim$ Exponential(λ^{α}) and sample inverse scale parameter for its normalized offspring intensity $\lambda_1^{\beta} \sim$ Gamma$(\alpha^{'}, \beta^{'})$.
3. For $n > 1$:
   (a) Sample $t_n > t_{n-1}$ from $\mathcal{PP}(\mu + \sum_{i=1}^{n-1} \alpha_i \beta_i (t - t_i))$
   (b) Sample point assignment $c_n \sim ddCRP(\eta, f, D)$. It indirectly determines cluster assignment and point types: $R(c_n)$, $R^{*}(c_n)$ and $R^{'}(c_n)$.
   (c) Sample hyper-cluster assignment $h_{R(c_n)} \sim CRP(\gamma)$, γ is the concentration parameter for CRP.
   (d) If $R(c_n)$ belongs to an existing hyper-cluster, then set $\alpha_n = \alpha_{h_{R(c_n)}}$ and $\beta_n = \beta_{h_{R(c_i)}}$. Otherwise, for a new hyper-cluster, sample its total offspring intensity $\alpha_h \sim$ Exponential(λ^{α}), and sample scale parameter for normalized offspring intensity $\lambda_h^{\beta} \sim$ InverseGamma$(\alpha^{'}, \beta^{'})$.

In the above, λ^{μ}, λ^{α}, $\alpha^{'}$, $\beta^{'}$ and k^{β} are hyper-parameters. It is worth noting that this hierarchical model extends the IBM in a similar way that the Chinese restaurant franchise (CRF) process (Teh et al. 2004) extends the CRP. The main difference is that the point clustering in our model is achieved via a ddCRP instead of a CRP with the consideration of branching structure. This hierarchial model can automatically discover the point clusters that share the same triggering scheme even when they are disjoint in spatiotemporal space.

Inference with Generic Metropolis-with-Gibbs Sampling

The purpose of the inference is to estimate the posteriors of branching structure, immigrant intensity and offspring intensity given observed points. Since it is not tractable analytically, we adopt the Markov chain Monte Carlo (MCMC) algorithm. Assume we have observed a set of points, $X = \{t_i\}_{i=1}^{N}$, for a time period $[0, T]$. We do not consider edge effect in this work, hence no point exists before time 0. As described in (Vere-Jones 1988), with the support of local Janossy density, the likelihood function for a realization X of a regular point process can be represented as:

$$L = \left(\prod_{i=1}^{N} \lambda(t_i)\right) \exp\left(-\int_0^T \lambda(t) dt\right), \qquad (3)$$

where $\lambda(t)$ denotes conditional intensity function. Unlike the traditional Hawkes process, the conditional intensity function in the IBM can be written separately for immigrants and offspring. Furthermore, directly modeling the branching structure grants us the computational simplicity to decompose the likelihood function into independent parts:

$$p(X|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, C) = p(I|\mu, C) \prod_{i=1}^{N} p(O_i | \alpha_{R(c_i)}, \beta_{R(c_i)}, C), \quad (4)$$

where I represents immigrants and O_i denotes the offspring whose parent is point i. The likelihood functions for I and O_i can be written as:

$$p(I|\cdot) = \frac{\prod_{t_i \in I} \mu(t_i)}{\exp\left(\int_0^T \mu(t) dt\right)}, \qquad (5)$$

$$p(O_i|\cdot) = \frac{\prod_{t_j \in O_i} \alpha_{R(c_i)} \beta_{R(c_i)} (t_j - t_i)}{\exp\left(\alpha_{R(c_i)} \int_{t_i}^T \beta_{R(c_i)} (t - t_i) dt\right)}. \qquad (6)$$

In some cases in which the conditional distributions of parameters are tractable, Gibbs sampling method can be used for inference. However, here we present a generic Metropolis-within-Gibbs algorithm (Neal 2000) despite the specific form of intensities. For Metropolis-within-Gibbs approach, each inference iteration updates parameters alternatively as Gibbs sampling does, while Metropolis-Hasting method is used for each parameter's update. In order to update a parameter w, a proposal distribution $q(\cdot)$ is used to generate a candidate value w^{*}. Its acceptance probability is defined as: $\min\left(1, \frac{q(w|w^{*})\tau(w^{*})}{q(w^{*}|w)\tau(w)}\right)$, where $\tau(\cdot)$ can be any un-normalized measure for parameter w. The second input of the $\min$ function is called Hastings ratio. In the following, we give the Hastings ratio for each parameter's update in the IBM:

$$A_{\mu} = \frac{p(\hat{\mu})}{p(\mu)} \prod_{t_i \in I} \left(\frac{\hat{\mu}(t_i)}{\mu(t_i)}\right) \exp\left(\int_0^T \mu(t) dt - \int_0^T \hat{\mu}(t) dt\right), \qquad (7)$$

$$A_{\alpha_{R(c_i)}} = \frac{p(\hat{\alpha}_{R(c_i)})}{p(\alpha_{R(c_i)})} \prod_{t_j \in R^{'}(c_i)} \left(\frac{\hat{\alpha}_{R(c_i)}}{\alpha_{R(c_i)}}\right) \cdot \exp\left(\sum_{t_j \in R^{'}(c_i)} \left(\alpha_{R(c_i)} - \hat{\alpha}_{R(c_i)}\right) B_{R(c_i)}\right), \qquad (8)$$

$$A_{\beta_{R(c_i)}} = \frac{p(\hat{\beta}_{R(c_i)})}{p(\beta_{R(c_i)})} \prod_{t_j \in R^{'}(c_i)} \left(\frac{\hat{\beta}_{R(c_i)}(t_j - t_{c_j})}{\beta_{R(c_i)}(t_j - t_{c_j})}\right) \cdot \exp\left(\sum_{t_j \in R^{'}(c_i)} \alpha_{R(c_i)} \mathbf{U}\right). \qquad (9)$$

For the above Hastings ratios, we assume the prior distributions of parameters are independent. In Eq. 8 and Eq. 9, intermediate variables are defined for notation simplicity: $B_{R(c_i)} = \int_{t_j}^{T} \beta_{R(c_i)}(t - t_j)dt$, $\hat{B}_{R(c_i)} = \int_{t_j}^{T} \hat{\beta}_{R(c_i)}(t - t_j)dt$, and $\mathbf{U} = B_{R(c_i)} - \hat{B}_{R(c_i)}$. Variables $\hat{\mu}$, $\hat{\alpha}_{R(c_i)}$ and $\hat{\beta}_{R(c_i)}$ indicate the candidate values drawn from proposal distributions, *e.g.*, Gaussian distribution. For updating the branching structure variables, the branching structure prior defined by Eq. 2 is used as the proposal distribution. The conditional prior and the proposal distribution cancel when calculating Hastings ratios, and only the likelihood ratio is left. There are three different cases for branching structure variable update: (1) update from immigrant to offspring. (2) update from offspring to immigrant, and (3) change parent. For the first case, the Hastings ratio can be represented as:

$$A_{c_i}^{I \to O} = \frac{\alpha_{R(\hat{c}_i)} \beta_{R(\hat{c}_i)}(t_i - t_{\hat{c}_i})}{\mu(t_i)} \cdot \prod_{t_j \in R'(c_i)} \mathbf{V} \cdot \exp\left(\sum_{t_j \in R'(c_i)} \mathbf{W} \right), \tag{10}$$

where $\mathbf{V} = \frac{\alpha_{R(\hat{c}_i)} \beta_{R(\hat{c}_i)}(t_j - t_{c_j})}{\alpha_{R(c_i)} \beta_{R(c_i)}(t_j - t_{c_j})}$ and $\mathbf{W} = \alpha_{R(c_i)} B_{R(c_i)} - \alpha_{R(\hat{c}_i)} B_{R(\hat{c}_i)}$ are intermediate variables for notation simplicity. The first part of Eq. 10 represents the likelihood ratio for point i, and the second part represents the likelihood ratio for all of its offspring indicated by $t_j \in R^{'}(c_i)$. Similarly, we can have the Hastings ratio for the second case:

$$A_{c_i}^{O \to I} = \frac{\mu(t_i)}{\alpha_{R(c_i)} \beta_{R(c_i)}(t_i - t_{c_i})} \cdot \prod_{t_j \in R'(\hat{c}_i)} \mathbf{V} \cdot \exp\left(\sum_{t_j \in R'(\hat{c}_i)} \mathbf{W} \right), \tag{11}$$

where $\mathbf{V}$ and $\mathbf{W}$ are as defined before. The second part of Eq. 11 also represents the likelihood ratio for all the offspring of point i, which are indicated by $t_j \in R^{'}(\hat{c}_i)$. For the third case, we have the Hastings ratio:

$$A_{c_i}^{O \to \hat{O}} = \frac{\alpha_{R(\hat{c}_i)} \beta_{R(\hat{c}_i)}(t_i - t_{\hat{c}_i})}{\alpha_{R(c_i)} \beta_{R(c_i)}(t_i - t_{c_i})} \prod_{t_j \in R'(c_i) \wedge t_j \in R'(\hat{c}_i)} \mathbf{V} \cdot \exp\left(\sum_{t_j \in R'(c_i) \wedge t_j \in R'(\hat{c}_i)} \mathbf{W} \right), \tag{12}$$

where $\mathbf{V}$ and $\mathbf{W}$ are as defined before. For the second part of Eq. 12, we use the notation $t_j \in R^{'}(c_i) \wedge t_j \in R^{'}(\hat{c}_i)$ to represent point i's offspring that change clusters when c_i is changed. Each of these Metropolis-Hasting updates can be performed several times before combining via Gibbs sampling. The Metropolis-within-Gibbs inference algorithm for the extended hierarchical IBM can be derived based on the above derivations with the consideration of hyper-cluster assignment.

Experiments

We conduct experiments on both synthetic and real-world data to evaluate the proposed IBM. The state-of-the-art approaches are compared to demonstrate its superiority.

Method	EMLL	MISD	BHawk	IBM(CRP)	IBM(Wind)
Diff	0.46	0.41	0.45	0.40	0.36
LogLik	-1063	-917	-1121	-862	-736

Table 1: Results of Diff and LogLik.

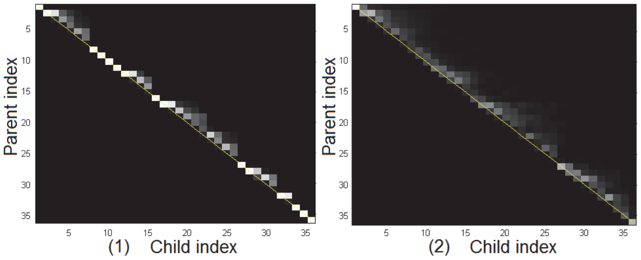

Figure 1: Estimated branching structure matrices.

Synthetic Data

In this section, we use the synthetic data generated from traditional Hawkes process to evaluate the IBM. Two triggering kernels, exponential and Weibull kernels, are used to generate the data. Immigrant intensities are set to 0.8 for both kernels. For each kernel, 130 synthetic temporal samples are generated on time interval $[0, 20]$. The simplified IBM with all points sharing the same offspring intensity is applied to the first 100 samples. Both the traditional CRP and the ddCRP with window decay function are adopted as branching structure prior for the IBM. Bayesian model averaging is applied to the estimated models obtained from the first 100 samples. The final model is used to (1) measure the difference between the true and estimated triggering kernels, and (2) measure the log-likelihood on the rest 30 samples. A relative distance called Diff defined by (Zhou, Zha, and Song 2013) is used to measure the difference between kernels. Three state-of-the-art approaches are compared with the proposed method: Hawkes process with expectation maximization on a lower bound of log-likelihood function (EMLL) (Yan et al. 2013), model independent stochastic declustering (MISD) (Marsan and Lengliné 2008) and Bayesian inference approach for Hawkes process (BHawk) (Rasmussen 2013). The comparison results for Diff and log-likelihood are given in Table 1. As we can see that the IBM can achieve the best performance for both Diff and log-likelihood. The ddCRP prior with window decay function outperforms the traditional CRP prior.

Besides, we select a synthetic sample from exponential kernel to visualize and demonstrate the IBM's performance on branching structure estimation. A matrix called branching structure matrix is used to demonstrate the estimation of branching structure. Fig. 1 (1) and Fig. 1 (2) show the branching structure matrices for the ddCRP prior and the CRP prior respectively. For these matrices, column indices represent child points and row indices represent parent points. The element in row i and column j represents the estimated probability of the parent-child relationship $c_j = i$. Bright color in the matrices indicates higher probability. As

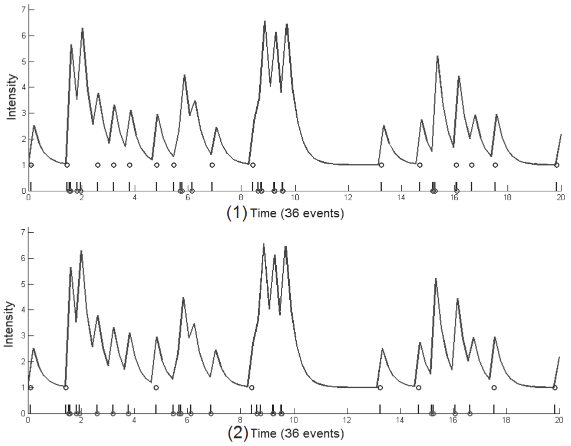

Figure 2: Estimated point types.

Method	HPP	SGCP	EMLL	MISD	BHawk	CPP	IBM
MSE	69.3	64.5	60.3	57.5	60.8	57.0	52.8
F1	-	-	0.70	0.75	0.71	0.76	0.79
SC	-	-	-	-	-	-	0.76

Table 2: Results of MSE and F1

we can see that both matrices show strong clustering behaviors. The ddCRP prior gives more clusters with fewer points in each cluster, while the CRP prior gives fewer clusters with more points in each cluster. Correspondingly, Fig. 2 (1) and Fig. 2 (2) show the results of point type estimation. In these figures, the vertical bars on time line denote the simulated points and the lines show the overall intensity. The circles on bars indicate that the points are estimated as offspring and the circles at higher positions indicate that the points are estimated as immigrants. As we can see in Fig. 2 (2), the CRP prior tends to underestimate the number of immigrants and exhibits strong ”rich gets richer” behavior. The ddCRP prior, shown in Fig. 2 (1), gives a more accurate estimation for the number of immigrants. Both priors can detect temporal clustering behaviors. The CRP prior tends to find coarser clusters, while the ddCRP prior tends to find finer clusters.

Real-world Application

For the real-world application, we apply our method to the water pipe failure prediction problem. Domain experts have observed that water pipe failures exhibit strong spatiotemporal clustering behaviors (Kleiner and Rajani 2001; Yan et al. 2013; Li et al. 2014; Lin et al. 2015; Li et al. 2015). One failure can cause other failures in adjacent spatiotemporal space. As a result, pipe failures can be categorized into two types: background failure that occurs due to material fatigue or corrosion, and triggered failure that is caused by another failure. It is desired for water utilities to accurately estimate both the type and amount of pipe failures.

In this experiment, we collected 922 failures from a metropolitan water supply network. The failures occurred in a district during 8 years. We treat each pipe failure as a point in spatiotemporal space. Hence, we can use our model for both failure type estimation and failure amount prediction. For the branching structure prior, we adopt a decay function considering both spatial and temporal distances:

$$f(d_S, d_T) = I(d_S \le a_S) \cdot I(d_T \le a_T) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{d_S^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{d_T}{\rho}\right), \tag{13}$$

where d_S and d_T represent spatial and temporal distances respectively. σ and ρ are pre-determined parameters. $I(\cdot)$ denotes a indicator function that returns 1 if the input condition satisfies and 0 otherwise. a_S and a_T are constants that determine the window sizes for spatial and temporal spaces. They can be set by domain experts as hard constraints to quickly filter out unrealistic branching structures. The hierarchical IBM is used for modeling the failures. It can automatically discover the failure clusters that share the similar failure triggering pattern.

In additional to EMLL, MISD and BHawk, homogeneous Poisson process (HPP), sigmoidal Gaussian Cox process (SGCP) (Adams, Murray, and MacKay 2009) and cascades of Poisson process (CPP) (Simma and Jordan 2010) are also compared with the IBM for failure amount prediction. We use 4, 5, 6 and 7 years data for training and the obtained models are used to predict the amount of the failures in the coming year. The mean square error (MSE) is used to measure the difference between the true and predicted failure amounts. For failure type categorization, the IBM, EMLL, MISD, BHawk and CPP are applied to all the failures to estimate their types. F1 score is used to measure the performances. Additionally, we use silhouette coefficient (SC) (Rousseeuw 1987), to measure the clustering performance for the hierarchical IBM's hyper-clusters. The spatial and temporal distances between the triggering and triggered failures are used to calculate silhouette coefficient. The other approaches do not have the ability to discover the hidden hierarchical structure. The results of MSE, F1 and SC are shown in Table 2. As we can see, the proposed method outperforms others for both MSE and F1 score and it can also achieve an accurate clustering on top of the failure clusters.

Conclusion and Future Directions

In this paper, we proposed the IBM, a Bayesian statistical model that generalizes and extends popular IPPs. It considers point intensity, interaction mechanism and branching structure simultaneously. The experimental results on both synthetic and real-world data demonstrate its superiority. There are also many potential venues for future work. It will be interesting to consider high order point interaction (Baddeley and Van Lieshout 1995), the connection between branching structure and causality measure of point processes (Kim et al. 2011) and the extension for multivariate IPPs.

Acknowledgement

National ICT Australia (NICTA) is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Adams, R. P.; Murray, I.; and MacKay, D. J. 2009. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *ICML*, 9–16. ACM.

Adams, R. P. 2009. *Kernel methods for nonparametric Bayesian inference of probability densities and point processes*. Ph.D. Dissertation.

Bacallado, S.; Favaro, S.; Trippa, L.; et al. 2013. Bayesian nonparametric analysis of reversible markov chains. *The Annals of Statistics* 41(2):870–896.

Baddeley, A. J., and Van Lieshout, M. 1995. Area-interaction point processes. *Annals of the Institute of Statistical Mathematics* 47(4):601–619.

Beal, M. J.; Ghahramani, Z.; and Rasmussen, C. E. 2001. The infinite hidden markov model. In *NIPS*, 577–584.

Blei, D. M., and Frazier, P. I. 2011. Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research* 12:2461–2488.

Diaconis, P., and Freedman, D. 1980. de finetti's theorem for markov chains. *The Annals of Probability* 115–130.

Diggle, P. J.; Fiksel, T.; Grabarnik, P.; Ogata, Y.; Stoyan, D.; and Tanemura, M. 1994. On parameter estimation for pairwise interaction point processes. *International Statistical Review* 99–117.

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

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

Hewlett, P. 2006. Clustering of order arrivals, price impact and trade path optimisation. In *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, 6–8.

Kim, S.; Putrino, D.; Ghosh, S.; and Brown, E. N. 2011. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology* 7(3):e1001110.

Kleiner, Y., and Rajani, B. 2001. Comprehensive review of structural deterioration of water mains: statistical models. *Urban water* 3(3):131–150.

Li, Z.; Zhang, B.; Wang, Y.; Chen, F.; Taib, R.; Whiffin, V.; and Wang, Y. 2014. Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. *Machine learning* 95(1):11–26.

Li, B.; Zhang, B.; Li, Z.; Wang, Y.; Chen, F.; and Vitanage, D. 2015. Prioritising water pipes for condition assessment with data analytics. *OzWater*.

Lin, P.; Zhang, B.; Wang, Y.; Li, Z.; Li, B.; Wang, Y.; and Chen, F. 2015. Data driven water pipe failure prediction: A bayesian nonparametric approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 193–202. New York, NY, USA: ACM.

MacEachern, S. N. 1999. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, 50–55.

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

Neal, R. M. 2000. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics* 9(2):249–265.

Neyman, J., and Scott, E. L. 1958. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–43.

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

Orbanz, P., and Roy, D. M. 2013. Bayesian models of graphs, arrays and other exchangeable random structures. *arXiv preprint arXiv:1312.7857*.

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

Ripley, B. D. 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)* 172–212.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20:53–65.

Simma, A., and Jordan, M. I. 2010. Modeling events with cascades of poisson processes. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, 546–555.

Snoek, J.; Zemel, R.; and Adams, R. P. 2013. A determinantal point process latent variable model for inhibition in neural spiking data. In *NIPS*, 1932–1940.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*.

Vere-Jones, D. 1988. An introduction to the theory of point processes. *Springer Ser. Statist., Springer, New York.*

Yan, J.; Wang, Y.; Zhou, K.; Huang, J.; Tian, C.; Zha, H.; and Dong, W. 2013. Towards effective prioritizing water pipe replacement and rehabilitation. 2931–2937.

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

Yu, H.; Xie, L.; and Sanner, S. 2015. The lifecyle of a youtube video: Phases, content and popularity. In *AAAI Conference on Web and Social Media*.

Zabell, S. L. 1995. Characterizing markov exchangeable sequences. *Journal of Theoretical Probability* 8(1):175–178.

Zhou, K.; Zha, H.; and Song, L. 2013. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*, 1301–1309.