Bayesian Nonparametric Approaches for Modelling Stochastic Temporal Events

Submitted by

Peng Lin

for the degree of

Doctor of Philosophy

School of Computer Science and Engineering

The University of New South Wales

October 2017

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Lin

First name: Peng Other name/s:

Abbreviation for degree as given in the University calendar: PhD

School: Computer Science and Engineering Faculty: Engineering

Title: Bayesian Nonparametric Approaches for Modelling Stochastic Temporal Events

Abstract 350 words maximum: (PLEASE TYPE)

Modelling stochastic temporal events is a classic machine learning problem that has drawn enormous research attentions over recent decades. Traditional approaches heavily focused on the parametric models that pre-specify model complexity. Comprehensive model comparison and selection are necessary to prevent over-fitting and under-fitting problems.

The recently developed Bayesian nonparametric learning framework provides an appealing alternative to traditional approaches. It can automatically learn the model complexity from data. In this thesis, I propose a set of Bayesian nonparametric approaches for stochastic temporal event modelling with the consideration of event similarity, interaction, occurrence time and emitted observation. Specifically, I tackle following three main challenges in the modelling.

1. Data sparsity. Data sparsity problem is common in many real-world temporal event modelling applications, e.g., water pipes failures prediction. A Bayesian nonparametric model that allows pipes with similar behaviour to share failure data is proposed to attain a more effective failure prediction. It is shown that flexible event clustering can help alleviate the data sparsity problem. The clustering process is fully data-driven and it does not require predefining the number of clusters.
2. Event interaction. Stochastic events can interact with each other over time. One event can cause or repel the occurrence of other events. An unexplored theoretical bridge is established between interaction point processes and distance dependent Chinese restaurant process. Hence an integrated model, namely infinite branching model, is developed to estimate point event intensity, interaction mechanism and branching structure simultaneously.
3. Event correlation. The stochastic temporal events are correlated not only between arrival times but also between observations. A novel unified Bayesian nonparametric model that generalizes Hidden Markov model and interaction point processes is constructed to exploit two types of underlying correlation in a well-integrated way rather than individually. The proposed model provides a comprehensive insight into the interaction mechanism and correlation between events.

At last, a future vision of Bayesian nonparametric research for stochastic temporal events is highlighted from both application and modelling perspectives.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

.... 12/10/2007

Signature Witness Signature Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed .....

Date 12/10/2017

To who I love.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Professor Fang Chen for her enormous support, timely guidance and constant encouragement. Fang is extremely supportive and caring to every student. Her personal charisma has deeply influenced me.

I am very much indebted to my co-supervisor Doctor Matt Zhang for being continuous sources of insightful guidance, valuable suggestions and cooperative friendship. Matt is an intelligent and inspiring mentor from whom I have learnt a lot about research and team work.

I would also like to thank all my friends and colleagues in the University of New South Wales and Data61. You all become part of my life.

Finally, I would like to thank my family for their selfless love and support in this long journey. Thanks for coming along with me.

Abstract

Modelling stochastic temporal events is a classic machine learning problem that has drawn enormous research attentions over recent decades. Traditional approaches heavily focused on the parametric models that pre-specify model complexity. Comprehensive model comparison and selection are necessary to prevent over-fitting and under-fitting problems.

The recently developed Bayesian nonparametric learning framework provides an appealing alternative to traditional approaches. It can automatically learn the model complexity from data. In this thesis, I propose a set of Bayesian nonparametric approaches for stochastic temporal event modelling with the consideration of event similarity, interaction, occurrence time and emitted observation. Specifically, I tackle following three main challenges in the modelling.

1. Data sparsity. Data sparsity problem is common in many real-world temporal event modelling applications,*e.g.*, water pipes failures prediction. A Bayesian nonparametric model that allows pipes with similar behaviour to share failure data is proposed to attain a more effective failure prediction. It is shown that flexible event clustering can help alleviate the data sparsity problem. The clustering process is fully data -driven and it does not require predefining the number of clusters. (This work has been published on [1])

2. Event interaction. Stochastic events can interact with each other over time. One event can cause or repel the occurrence of other events. An unexplored theoretical bridge is established between interaction point processes and distance dependent Chinese restaurant process. Hence an integrated model, namely infinite branching model, is developed to estimate point event intensity, interaction mechanism and branching structure simultaneously. (This work has been published on [2])

3. Event correlation. The stochastic temporal events are correlated not only between arrival times but also between observations. A novel unified Bayesian nonparametric model that generalizes Hidden Markov model and interaction point processes is constructed to exploit two types of underlying correlation in a well-integrated way rather than individually. The proposed model provides a comprehensive insight into the interaction mechanism and correlation between events. (This work has been published on [3])

At last, a future vision of Bayesian nonparametric research for stochastic temporal events is highlighted from both application and modelling perspectives.

Contents

List of Figures

List of Tables

List of Abbreviations

ABBREVIATIONS	FULL EXPERSSIONS
BNP	Bayesian nonparametric
DP	Dirichlet process
HDP	Hierarchical Dirichlet process
DPMM	Dirichlet process mixture model
MSE	Mean square error
MCMC	Markov chain Monte Carlo
AUC	Area under the curve
EM	Expectationmaximization
DPMHBP	Dirichlet process mixture of hierarchical beta process
CWM	Critical water main
RWM	Reticulation water main
BeP	Bernoulli process
BP	Beta process
HBP	Hierarchical beta process
PVC	Polyvinyl chloride
SVM	Support vector machine

IBM	Infinite branching model
IPP	Interaction point process
DDP	Dependent Dirichlet process
bPP	Basis point process
ddCRP	Distance dependent Chinese restaurant process
$\mathcal{PP}$	Poisson process
IRM	Infinite relational model
CRF	Chinese restaurant franchise
EMLL	Hawkes process with EM on a lower bound of log-likelihood function
MISD	model independent stochastic declustering
BHawk	Bayesian inference approach for Hawkes process
HPP	Homogeneous Poisson process
SGCP	Sigmoidal Gaussian Cox process
CPP	Cascades Poisson process
SC	Silhouette coefficient
iHSMM-IPP	Infinite hidden semi-Markov modulated interaction point process model
SSM	State space model
HMM	Hidden Markov model
HSMM	Hidden semi-Markov model
PMCMC	Particle Markov chain Monte Carlo
SMC	Sequential Monte Carlo
PG	Particle Gibbs
REDD	Reference energy disaggregation dataset

Chapter 1

Introduction

1.1 Objective and Challenges

The evolving of our world can be regarded as series of stochastic temporal events. The research of stochastic temporal events modelling has drawn enormous attention during the past few decades. It has wide applications in various areas, such as financial modelling, social event analysis, infrastructure failure prediction, seismological and epidemiological forecasting.

In general, the objective of stochastic temporal events modelling is to learn the mechanism of event generation from observed historical events and then apply it to forecast forthcoming new events characteristics, *e.g.*, occurrence time, latent state, observable appearance. Fig. 1.1 gives an illustration of temporal events characteristics and generation mechanisms.

The characteristics of a temporal event consist of an occurrence time t, a latent state s and an observable appearance y. The generation mechanisms describe the rules of generating events characteristics. Occurrence time of an event is usually impacted by the occurrence times of its predecessors. In Fig. 1.1, the red arrow curves indicate the mechanism of occurrence time generation. It is worth noting that,

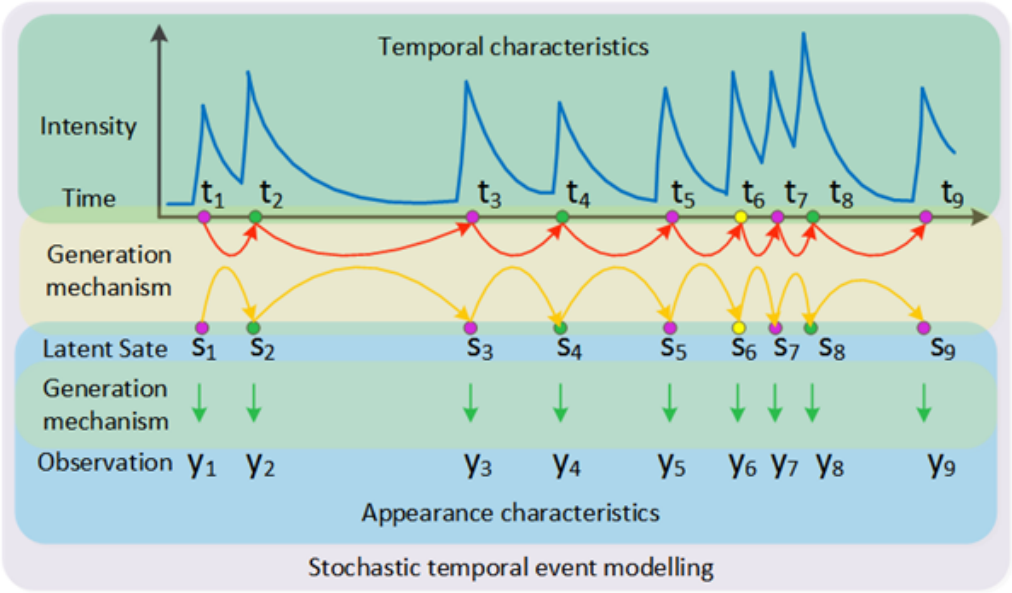

Figure 1.1: Characteristics and generation mechanisms of temporal events.

as an illustration, it only shows the first order of occurrence time relationship, *i.e.*, an events time is only impacted by the previous event's occurrence time. Similarly, the latent state of an event is influenced by the previous event's state. The yellow arrow curves show the generation mechanism of latent states.

Fig. 1.1 also illustrates an example possessing Markov property, i.e., an event's latent state is only determined by the state of its previous event. The generation of event appearance is usually governed by its latent state, as illustrated by green arrow lines. Indeed, the real-world scenarios are much more complicated than the illustrated situation whose purpose is for introducing the components of the modelling problem only.

Concrete examples can help us better understand the problem. Using a Twitter user's tweets stream as an example, a tweet has a time, a latent topic and the content. The occurrence time of a new tweet is impacted by the users previous tweets. Its topic is influenced by previous tweets topics. The content is generated based on the topic. Another example can be a market trading event stream. A trading event

has its occurrence time which can put an impact on the following events, and a latent trading intention which governs the observable prices and amount. In the area of infrastructure asset predictive maintenance, an infrastructure asset failure, *e.g.*, drinking water pipe burst, has a failure time, failure type, e.g., circular break, and failure cost, e.g., repair cost.

The main challenges of stochastic event modelling come from two aspects: the intrinsic complexity of stochastic events and the limitation of traditional modelling approaches. On the one hand, real-world stochastic events are complicated. In order to fully understand them, people need to consider as many aspects of events as possible, *e.g.*, occurrence time, latent state, emitted observation, the correlation between events, generation or triggering mechanism. These aspects often need to be considered together, which makes the problem extremely difficult. Determining model complexity beforehand is almost impossible. On the other hand, most traditional approaches of stochastic event modelling are parametric, which means the form of the model needs to be pre-defined beforehand based on somewhat biased assumptions or priors. A model is competent only when its prior can adequately capture the true underlying data structure while an invalid prior makes the model vulnerable to over-fitting or under-fitting problems. The support of a comprehensive model selection process is often required for finding the proper model form. The process is computationally costly and cannot guarantee the optimal solution. Hence, traditional approaches are deficient in flexibility to model complex stochastic events.

Bayesian nonparametric (BNP) approaches have attracted increasing interests in recent years due to its flexibility and expression power for modelling complicated real-world scenarios. A BNP approach fits a single model whose complexity is determined by data rather than comprehensive model selection over a family of parameterized models with predefined different complexities. Comparing with traditional parametric approaches, the number of parameters in a BNP model can increase with data. Therefore, its model form adapts to the data. A BNP approach

has the ability to model over infinite dimensional function or measure spaces. It supplies a broad class of flexible priors allowing data to speak for itself. Hence, it have been widely used for modelling various data structures, *e.g.*, array, partition, matrix, tree, network, graph and temporal sequence.

Another benefit of BNP approaches is that many existing Bayesian parametric approaches can be enhanced by introducing a nonparametric prior that incorporates more data information which the original model cannot consider. The derived nonparametric model will be more expressive, hence more powerful. As a result, BNP approaches open another door for us to better model stochastic temporal events.

The thesis focuses on Bayesian nonparametric-based stochastic temporal events modelling. I aim to demonstrate how to use nonparametric theory building flexible but principled models to understand interaction and correlation between stochastic events and predict the incoming events, thereby helping solve real-world problems.

1.2 Thesis Overview

A fundamental assumption of the thesis is that there exist learnable patterns(*e.g.*, analogy, correlation and interaction among events) hidden in the real-world temporal events, despite how complicated they are. Understanding these patterns will help yield effective solutions. Although the underlying true physical relationship patterns of stochastic events are generally complicated, they can often be modelled as a combination of infinite multiple simpler patterns, with the support of BNP approaches.

Dirichlet process (DP)[4] and its variants (see Sec. 2.2) provide elegant tools for exploring such complicated relationships. As one of the most popular BNP approaches, it has been widely adopted as a flexible BNP prior over countably-infinite partitions of a space. The realization of a DP is discrete, which means DP is not suitable for directly modelling continuous variables. However, it fits perfectly

for the cases of modelling an unknown number of relationships among patterns via combining forces of possibly infinite statistical models. The literature is rich on the subjects of how DP model family (abbreviated to DPs) are utilized to extend existing models to represent complicated underlying patterns of data and outperform traditional approaches (see survey in Sec. 2.2.7)

A particularly important and foundational model derived from Dirichlet process is *Dirichlet process mixture model (DPMM)* [5] (see Sec. 2.2.4) It has been the cornerstone of many BNP approaches. In DPMM, the observed data are assumed to belong to one of an infinite number of clusters and the data from the same cluster share the same distribution which is distributed according to a random measure.

In this thesis, with the supports of DPs and DPMM, I investigate the problems in modelling stochastic temporal events from the following three perspectives:

Conducting flexible grouping to facilitate sparse event prediction Many real-world stochastic events exhibit sparseness feature. Such sparse ness brings significant challenges to the modelling. For example, in the scenario of water pipe failure prediction, the data sparsity problem makes the prediction model difficult to train as very few pipes have failure records during the observation period. In the thesis, with the support of DPMM, a hierarchical BNP model that clusters similar pipes together to share failure events is developed to conduct more accurate and efficient failure event prediction (see Chapter 3 for details).

Learning infinite branching structure to generalize interactive point process Many stochastic events series such as stock trades, earthquakes and epidemics, usually exhibit strong interactive patterns and cluster properties in both temporal space and feature space. In other words, one event can trigger the occurrences of others. Interaction point processes (IPPs) (see Sec. 2.4) represent a class of stochastic point processes that can model the interaction between points. In this thesis, a variant of DP, *distance dependent Chinese restaurant process(ddCRP)*

(see Sec. 2.2.6), is adopted to generalize and improve the IPPs, yielding a Bayesian nonparametric branching model(See Chapter 4 for further explanation).

Modelling infinite latent states to capture observation and arrival time correlation In stochastic events series, the correlation exists not only between events' emitted observations, but also between their arrival times. State space models (*e.g., hidden Markov model*) and stochastic interaction point process models (e.g., Hawkes process) have been studied extensively yet separately for the two types of correlations. In this thesis, *hierarchical Dirichlet process(HDP)*(see Sec. 2.2.5) is adopted to model the state transition matrix to construct a Bayesian nonparametric model (details are discussed in Chapter 5) that considers both types of correlations.

1.3 Thesis Organization

We organize the rest of thesis as follows: Chapter 2 Background, Chapter 3 Bayesian nonparametric approach for sparse event prediction, Chapter 4 Infinite Branching Model, Chapter 5 Infinite Hidden Semi-Markov Modulated Interaction Point Process, Chapter 6 Conclusions and future works.

Chapter 2: Background

In Chapter 2, I first review the background of BNP approaches. Then, I provide an overview of the stochastic events models investigated in this thesis. Also, I introduce the foundation of Markov chain Monte Carlo (MCMC) inference framework.

Chapter 3: Bayesian Nonparametric Approach for Sparse Event

In chapter 3, I present a novel BNP approach, namely Dirichlet process mixture of hierarchical beta process model, for sparse temporal event prediction and apply it to the task of water pipe failure prediction. A prediction model that can predict future failure probability is developed and hence the high-risk pipes can be prioritized for preventative physical condition assessment. As a result, disastrous infrastructure failures can be prevented proactively.

Dealing With Sparse Event Using Flexible Grouping Like many other real-world machine learning applications, water pipe failure events prediction encounters the sparse data problem, as very few pipes have failure records during the observation period. Such sparsity makes traditional methods incompetent for accurate pipe failure prediction since most pipes do not have failure data for training. I propose tackling this sparse data problem by sharing failure data via a flexible hierarchical modelling of failure behaviours. The key concept is a flexible grouping scheme that clusters pipes with similar failure pattern together for modelling so that failure data can be shared by similar pipes for training. The failure probability of a pipe segment is modelled by *beta process* and the failure records of pipes are represent with an infinite binary matrix. The pipes with similar failure patterns are presumed to constitute a cluster whose pipes' average failure probability is distributed based on a new beta process. The Dirichlet process is adopted as a flexible prior for the pipe-cluster assignment variable with no assumptions on the number of clusters. Based on this tactic, the Dirichlet process mixture of hierarchical beta process (DPMHBP) model is constructed.

Model Inference The proposed model has no analytical solution. I develop an approximated yet computational efficient Metropolis-within-Gibbs sampling method

for model parameter inference. Also, a Gibbs sampling inference step for large-scale datasets is derived by making use of the sparsity property of failure records to obtain an approximated conjugacy.

Failure Prediction The proposed model is applied to a metropolitan water supply network. Area under curve (AUC) is calculated for measuring the performances of different approaches and one-sided paired t-test is performed on AUC to evaluate the significance of performance difference. The comparison results show that the proposed approach significantly outperforms the state-of-the-art prediction methods, including Weibull[6], Cox [7], SVM, HBP [8] . Many failures could be prevented and significant economic and social savings could be brought to the water utility if the proposed method were applied.

Chapter 4: Bayesian Nonparametric Approach for Event Interaction

In chapter 4, I propose the infinite branching model (IBM), a Bayesian nonparametric model that generalizes and extends some popular interaction point processes(IPPs). IBM redefines the IPP as an infinite mixture of basis point processes with the aid of a distance dependent prior over infinite branching structure that describes the relationships between points.

Modelling the Infinite Branching Structure by Point Connection Many IPPs, *e.g.*, Hawkes process, can be redefined equivalently as Poisson cluster processes which are constituted by collection of basic Poisson processes following a specific branching structure that describes the relationship between events. This branching structure is defined in temporal space and effective to capture the impact of event interaction on arrival time. However, the strengths of original IPPs are insufficient in terms of capturing the branching structure in observation(feature

space) that reflects the interactive and cluster trait as well. To fill this gap, I resort to *distance dependent Chinese restaurant process(ddCRP)* [9] (see Sec. 2.2.3), which defines a class of non-exchangeable distributions over branching structure. In dd-CRP, observations constitute infinite number of branching by connecting each other based on the distance between them. The distance metrics can be defined on time, physic or feature spaces. DdCRP is placed as an infinite prior over the branching structure, in which the interaction between points is depicted by their connection. Then a Bayesian nonparametric model, namely Infinite Branching Model (IBP) is formed. IBM can learn the point events intensity, interaction mechanism and branching structure simultaneously. The cluster traits in both temporal space and feature spaces are captured in an integrated way. Unlike traditional IPPs where the offsprings share the same intensity, IBM allows different offspring intensities for different clusters, which grants more flexibility for modelling real-world events. In addition, I construct hierarchical IBM model in which similar point clusters form a hyper-cluster sharing the same offspring intensity. Hierarchical IBM extends IBM model in a similar way that the Chinese restaurant franchise (CRF) process [10] extends the CRP. It can automatically discover the point clusters that share the same triggering scheme even when they are disjoint in spatiotemporal space.

Inference Because the proposed model is not tractable analytically, a generic Metropolis-within-Gibbs sampling method is developed for model parameter inference. Due to the cluster trait of IPPs, the immigrant term and offspring term in likelihood function are independent conditioning on the latent branching structure. Therefore, calculation of Hastings ratios for latent branching parameters update can be simplified by considering only three distinct cases.

Empirical Study Experiments are constructed on both synthetic and real-world data to evaluate the proposed model. I firstly demonstrate the IBM's performance on branching structure estimation based on synthetic data that generated from tradi-

tional Hawkes process with two triggering kernels: exponential and Weibull kernels, respectively. For the real-world application, the proposed method is applied to the water pipe failure prediction problem. Hierarchical IBM is compared with methods such as Model independent stochastic declustering (MISD) [11], Bayesian Hawkes process(BHawk) [12], homogeneous Poisson process (HPP), sigmoidal Gaussian Cox process (SGCP)[13] and cascades of Poisson process (CPP)[14], etc. For failure amount prediction. The comparison result shows that the proposed method outperforms others for accurate failure clustering. The superiority relies on the model's capability to capture the event interaction with hierarchical structure exhibited in spatiotemporal space.

Chapter 5: Bayesian Nonparametric Approach for Event Correlation

In this chapter, I propose Infinite Hidden Semi-Markov Modulated Interaction Point Process(iHSMM-IPP) model to investigate stochastic events considering the observation correlation and arrival time correlation in a unified manner.

Exploring Arrival Time Correlation and Observation Correlation Simultaneously In stochastic events series, the correlation exists not only between events emitted observations but also their arrival times. Hidden Markov model (HMM) [15] has been a powerful tool for modelling the correlation between observations in the way that the latent state behind an event observation is influenced by its predecessors. However, HMM does not considering the correlation between arrival times. Interaction point process (IPP) is widely adopted for modelling arrival time correlation by defining a conditional intensity that depicts the interaction that an event arrival time depends on all the previous events. However, it lacks of the capability of modelling events latent states and their interactions. Inspired by hidden semi-Markov model (HSMM) [16, 17] that allows each state to emit a

sequence of observations, a novel Bayesian nonparametric model, iHSMM-IPP, is proposed to acquire the merits of both HMM and IPPs to model the two types of correlation simultaneously with an integrated manner. The core of proposed model is a latent semi-Markov state chain with infinitely countable number of states which govern both the observation emission and new event triggering mechanism. Hierarchical Dirichlet process(HDP) is employed as the prior over infinite latent state transformation matrices. The resulting model unifies and generalizes HMMs and IPPs and can model stochastic events series by simultaneously considering the correlations between arrival times and between emitted observations. As a Bayesian nonparametric model, it can infer the number of states based on events data.

Inference The proposed iHSMM-IPP model faces challenges for posterior inference: strong correlation nature of its temporal dynamics and non-Markovianity introduced by the event triggering mechanism. As traditional sampling methods for high dimensional probability distributions, e.g., MCMC, sequential Monte Carlo (SMC), are unreliable when highly correlated variables are updated independently, I develop the inference algorithm within the framework of particle MCMC (PMCMC) [18], a family of inferential methods that use SMC to construct a proposal kernel for an MCMC sampler. For tackling the non-Markovianity, ancestor resampling scheme [19] is incorporated into the inference algorithm, which uses backward sampling to improve the mixing of PMCMC and thereby provides effective sampling.

Empirical Study The superiority of the proposed model is demonstrated by synthetic data experiment and two real-world data applications. For the synthetic data experiment, the synthetic data is drawn via Gaussian emission HMM and several related methods are compared including sticky HDP-HMM [20], HDP-HSMM [21] and marked Hawkes process [12]. The first real-world application is to understand energy consumption behaviours of households based on reference energy disaggregation dataset. The appliance types can be modelled as latent states in the proposed

iHSMM-IPP model and the readings are state's emitted observations governed by Gaussian distributions. The triggering kernels of states in the model depict the influences of appliances on triggering the following energy consumption. The second application is to understand water pipe networks behaviours and impact. The failure types are modelled as latent states and labour hours for repair are modelled as states' emissions, which are Gaussian-distributed. The proposed model outperforms the other methods in both applications due to the fact that it well utilizes both the observed information and occurrence times while others only consider part of the information or have limitations on model flexibility.

Chapter 6: Conclusions and Future Work

In Chapter 6, I conclude by summarizing the contributions of this thesis. I also summarize potential directions of future work.

Chapter 2

Background

In this chapter, the techniques that are related to the thesis are reviewed. The review mainly focuses on three parts. First, the related Bayesian nonparametric approaches that are utilized in the thesis for improving traditional stochastic temporal event modelling are summarized. Second, I review some of the most popular frameworks that have been widely adopted for temporal event modelling, particularly for modelling event interaction and event latent state. At last, the inference methods that are used for inferring the parameters of the proposed models are introduced. For each technique, I briefly introduce its theory and summarize its applications.

At the beginning, I introduce Dirichlet process (Sec. 2.2) and its extensions. Then I review beta process (Sec. 2.3) which depicts the occurrence of temporal series as a sequence of independent binary variable. Next, the theory of interaction point process is introduced in Sec. 2.4. After that, a quick review of hidden Markov model and its extension is given in Sec. 2.5. At the end, as exact Bayesian inference is infeasible for the proposed model, Markov chain Monte Carlo methods are used. They are introduced in Sec. 2.6.

2.1 Bayesian Nonparametric: from Finite to Infinite

Classic Bayesian approach derives the posterior distribution based on both likelihood and prior:

$$p(\theta|X) \propto p(\theta) \times p(X|\theta), \tag{2.1}$$

where, likelihood $p(X|\theta)$ defines a family of probability distributions over observations X with parameter θ restricted in a finite-dimensional space. A density function $p(\theta)$ is placed to represent the prior beliefs over the parameter. When we infer the posterior of parameter θ given observations X, the dimension of parameter is fixed, so the complexity and scale of methods are fixed.

Nonparametric methods have achieved remarkable success in frequentist statistics (non-Bayesian) [22]. This kind of methods makes fewer assumptions about the form of probability distributions and the complexity of models can be determined from data.

Bayesian nonparametric (BNP) inherits the schemes of traditional parametric Bayesian and the concept of frequentist nonparametric. Distinctive from classic Bayesian, a BNP approach is built through more flexible and expressive parameter that is designed as a general stochastic process, an infinite-dimensional random variable. The inference process can be formulated [23] as:

$$p(G|X) \propto p(G) \times p(X|G), \tag{2.2}$$

where, likelihood $p(X|G)$ represents a far richer family of distributions over X with parameter which lies in infinite-dimensional space (G denotes a stochastic process). The prior believe $p(G)$ represents a probability measure on infinite-dimensional variables. In the inference process, the parameter complexity adapts and fits to the data automatically.

Using probability measure over a general stochastic process $p(G)$ rather than probability distribution over fix-dimensional parameter $p(\theta)$ as the prior grants BNP model infinite flexibility and expressiveness. Among all the models, Dirichlet process [4] and beta process [24] are extensively adopted and have become the pillars for building many sophisticated Bayesian non-parametric models.

2.2 Dirichlet Process and Its Variants

Dirichlet process (DP) [4] is one of the most popular Bayesian non-parametric processes. It has been applied with tremendous success in diverse domains, such as computer vision [25, 26], musical analysis [27, 28], social network analysis [29], natural language parsing [30] and information retrieval [31], etc. In this section, we will make a quick review of current Dirichlet process models and its extensions.

2.2.1 Dirichlet Process

Dirichlet Process defines a random measure over a family of probability distributions and usually serves as a prior over random partitions. A Bayesian nonparametric model with DP prior does not set any assumptions on the number of partitions. Instead, it allows the number to grow as the data observation increases.[5]

Suppose that G is a probability distribution over a measurable space Θ. If the marginal distribution of G is DP, which is parameterized by a concentration parameter α and a base measure H, then for any finite measurable partition $T_1, T_2, ..., T_K$ of Θ, random vector $(G(T_1), G(T_2), ..., G(T_K))$ will obey Dirichlet distribution.

$$(G(T_1), G(T_2), ..., G(T_K)) \sim Dirichlet(\alpha H(T_1), \alpha H(T_2), ..., \alpha H(T_K)). \tag{2.3}$$

In other words, the probabilities that G with any finite partition of Θ obeys a Dirichlet distribution.

The two parameters α and H can be intuitively explained. Base measure H is the mean of DP: $E(G) = H$. On the other side, the concentration parameter can be understood as an inverse of variance: $V(G) = \frac{H(1-H)}{1+\alpha}$.

Let $\theta_1, \theta_2, ..., \theta_n$ denote a sequence of independent samples from G. Posterior over G is also a DP.

$$G|\theta_1, ..., \theta_n \sim DP(\alpha + n, G_n), \tag{2.4}$$

where

$$G_n = \frac{1}{\alpha + n}(\alpha H(\theta) + \sum_{i=1}^{N} \delta(\theta = \theta_i)). \tag{2.5}$$

Thus, DP is the conjugate prior for arbitrary distribution over a measurable space Θ.

The posterior shows a weighted average of prior base measure and the empirical measure. If the weighted factor $\alpha \to 0$, the prior becomes non-effective and the posterior distribution will only be given by the empirical distribution. On the other hand, if the observations is sufficiently large, such that $n \gg \alpha$, the posterior will be dominated by the empirical distribution. This property makes DP suitable to estimate the true underlying distribution.

Suppose θ_{n+1} is a new sample of G, using the conjugacy property, the predictive distribution over θ_{n+1} can be obtained directly as follows.

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{1}{a + n}(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}(A)). \tag{2.6}$$

Dirichlet process describes essentially a "distribution over distribution". Several constructive representations from different schemes are proposed. Stick-breaking construction and Chinese restaurant process are the most popular two of them.

2.2.2 Stick-breaking Construction of DP

Stick-breaking process [32] generates an infinite sequence of independent random variables as follows.

$$\begin{aligned} V_i &\sim Beta(1, \alpha) \\ \pi_i &= V_i \prod_{j=1}^{j-1} (1 - V_j). \end{aligned} \tag{2.7}$$

When π_i is drawn this way, we denote $\pi_i \sim GEM(\alpha)$. It is clear that $\sum_{k=1}^{\infty} \pi_k = 1$. Based on stick-breaking process, a fundamentally important construction process of DP can be obtained as follows [32].

Consider an arbitrary measurable space ω and a probability measure H on ω. Any $G \sim DP(\alpha, H)$ can be formulated as

$$\begin{aligned} \theta_k &\sim H \\ G(\theta) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta), \end{aligned} \tag{2.8}$$

where, δ_{θ_k} denotes a unit mass at point θ_k. Clearly G is also a measure and is composed of a weighted sum of infinite point masses, therefore a draw from G is discrete with probability one.

Because of the discreteness, the sample θ_i from DP will be repeated during the generation process, a clustering property is thereby manifested. Such clustering property is explicit in another construction process of DP, Chinese restaurant process (see Sec. 2.2.3).

Stick-breaking representation plays an important role in Bayesian nonparametric. It has been the cornerstone of constructive definitions of many nonparametric models. For instance, it is employed to generate extensions of DP that allow dependence across a collection of distributions (see Sec. 2.2.7).

2.2.3 Chinese Restaurant Process

Chinese restaurant process(CRP) is another widely-used construction of DP. It describes the marginal distribution over a random partition. Specifically, if the prior on random variable G is a Dirichlet process, then the CRP defines how observations are assigned to clusters when we integrate out G [4].

Chinese restaurant process exhibits the clustering property of DP in an explicit manner via a metaphor. Imagine there is a Chinese restaurant that has an infinite number of tables that correspond to clusters. A sequence of customers that correspond to data points enter and select a table to sit. The first customer sits at the first table. The succedent customers sit at a previously occupied table with probability proportional to the number of customers already sitting at the table, and they sit at an unoccupied table with probability proportional to a concentration parameter:

$$p(z_l = r|z_{-l}, \alpha) \propto \begin{cases} \frac{n_r}{n-1+\alpha} & \text{if } r \leqslant k \\ \frac{\alpha}{n-1+\alpha} & \text{if } r = k+1, \end{cases} \tag{2.9}$$

where, z_l indicates a customer, z_{-l} denotes all the customers that enter the restaurant before z_l, r indicates a table index, and k represents the current sited number of tables. n_r is the amount of customers sitting on table r and α is the concentration parameter for CRP, controlling the probability that a new customer selects an unoccupied table to sit. This metaphor has turned out to be useful in considering various generalizations of the Dirichlet process.

The CRP offers an exchangeable distribution over the table assignments z_l. The joint distribution is invariant to the order of customers. The procedure of assigning a table for a customer can be performed as he or she is the last customer entering the restaurant. As described by Eq. 2.9, the i-th customer sits at an occupied table with a probability proportional to the number of customers who are already sitting at that table. He or she sits at an unoccupied table with a probability proportional to the concentration parameter α. As currently occupied tables are more likely to get

new customers, CRP possesses a kind of rich get richer property. Theoretically, the number of occupied tables K will almost surely grow logarithmically with dataset size [33].

As a prior over partition of the data, CRP is exchangeable, by which it means the probability that a table is selected only depends on the number of pre-existing customer at the table. This exchange-ability can be exhibited from the fact that the joint distribution for $(\theta_1, ..., \theta_n)$ is invariant to order, which can be obtained easily via Eq. 2.9.

2.2.4 DP Mixture Model

DP can be adopted as the prior of latent parameters of mixture model, yielding the Dirichlet process mixture model (DPMM) [5] which comprises a countable infinite number of mixture components and self-adjust the number of mixture for fitting observed data.

Suppose we have n observations, denoted by $x_i (i \in [1 : n])$. The generative process for the DP mixture model is as follows:

$$\begin{aligned} G|\alpha, H &\sim DP(\alpha, H) \\ \theta_i &\sim G \\ x_i|\theta_i &\sim F(\theta_i), \end{aligned} \tag{2.10}$$

where, $F()$ is a likelihood function parameterized by θ_i. Due to the clustering property of DP, some $\theta_i (i \in [1 : n])$ take the same value and the observations $x_{i:n}$ that share the same parameter θ_i belong to the same cluster. Each data point x_i is drawn from a component of the mixture model. z_i is the component indicator for x_i. θ_k represents the parameter for component k. Fig. 2.1 shows the graphic model of DPMM.

In practical application, the clusters number in DPMM can be automatically

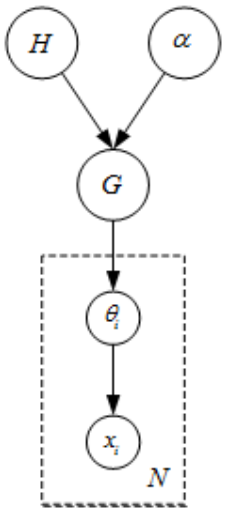

Figure 2.1: DP mixture model

inferred from data by using Bayesian posterior inference framework. Literature [34] makes a comprehensive review of MCMC inference methods for DP mixture model.

2.2.5 Hierarchical Dirichlet Process

Hierarchical Dirichlet process (HDP) [10] is a distribution over a group of random probability measures. It defines a set of random probability measures G_j, one for each group, and a global random probability measure G_0. The global measure G_0 is distributed as a Dirichlet process with concentration parameter γ and base probability measure H. The random measures G_j are distributed as a Dirichlet process with the base probability measure G_0.

$$\begin{aligned} G_0|\gamma, H &\sim DP(\alpha, H) \\ G_j|\alpha_j, G_0 &\sim G_0. \end{aligned} \tag{2.11}$$

Akin to Dirichlet Process, HDP can be constructed using a metaphor of Chinese restaurant franchise process [10], which extends Chinese restaurant process by al-

lowing multiple restaurants (corresponding groups) share the dishes (corresponding clusters) from a global menu.

Hierarchical Dirichlet process yields a natural way of modelling groups of data where the same clusters are shared among all the groups. Assume we have J groups of data. For each of group j, the observations x_{ji} $(i \in [1, 2, ...])$ are drawn from the model $F(\theta_{ji})$ with parameters θ_{ji} be i.i.d. random variables distributed as G_j that is drawn from $DP(\alpha_j, G_0)$ which itself is drawn from another Dirichlet process$DP(\gamma, H)$. This produces the HDP mixture model, the generative formulation is represented as

$$\begin{aligned} \theta_{ji}|G_j &\sim G_j \\ x_{ji}|\theta_{ji} &\sim F(\theta_{ji}). \end{aligned} \tag{2.12}$$

In HDP mixture model, the HDP supplies a prior for the hierarchical partition of the data. Not only observations within one group x_{ji_1}, x_{ji_2} share the atom (parameters θ), but also observations across groups $x_{j_1i_1}, x_{j_2i_2}$ might share the atom.

2.2.6 Distance Dependent Chinese Restaurant Process

Distance dependent Chinese restaurant process(dd-CRP)[9] is a generalization of the Chinese restaurant process (CRP) that is an exchangeable prior over 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 supplies a flexible class of distributions over partitions that allow for dependencies between observations.

The generative process of dd-CRP is based on Chinese restaurant metaphor, where a sequence of customers (correspond to data points) enter and select a table (correspond to clusters) to sit. However, unlike traditional CRP, dd-CRP represents with customer assignments instead of table assignments. Specifically, dd-CRP draws

the customers linkages based on the customers distance measurement. A customer is either linked to another customer with probability proportional to a decay function output depending on their distance or self-linked with probability proportional to a concentration parameter. The customers who are linked together will sit on the same table thus belongs to the same cluster while the customers who self-linked will sit on a new unoccupied table. The customer assignment is based on

$$p(c_i = j|\alpha, f, D) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{if } i = j, \end{cases} \tag{2.13}$$

where, d_{ij} is the distance between customers i and j. $f()$ is decay function, which defines how distance measurement impact the linkage probability. Decay function is a non-increasing and finite non-negative, and satisfies $f(\infty) = 0$.

In dd-CRP, table assignment can be obtained via customer assignment as a byproduct. It is worth noting that a same table assignment can correspond to several different customer assignments.

The dd-CRP defines a probability distribution over partitions using the connection among observations. In this setting, the connection can be transmitted from one observation to another. This property allows it to model a complicated scenario where influence between observations can be transmitted from one to another.

2.2.7 Other Variants of DP

The research on Dirichlet process has attracted a fair amount of attention. In Bayesian non-parametric framework, DP mainly serves as a flexible prior on partition over a measurable space, allowing to learn without predefining the number of partition, such as [35], [36], [37],[38], etc.

DPs have been adopted to reinvent some classic machine learning model as the nonparametric version models with enhanced flexibility and expressiveness. Such as

infinite hidden Markov model [39], infinite support vector machine [40], nonparametric k-means [41], Dirichlet process mixtures of generalized linear models[42], etc.

The models derived from Dirichlet process have a common character: the observations are assumed infinitely exchangeable, which means the order of observations does not affect the joint distribution of data. However, in many circumstances, this assumption does not hold. For example, in the hierarchical clusters structure, the data from different groups may not be exchangeable. In the time series analysis, the observations are often correlated in time space. Another example that we have studied in-depth in this thesis is that in stochastic events series such as water pipe failures, the occurrence of one failure may trigger another in the adjacent spatial-temporal space.

A series of studies introduce hierarchical or nested structure to DP to model multilevel cluster structure. A pioneering work is Hierarchical Dirichlet Process (HDP) [10]. As a two-layer model, the base measure for the child Dirichlet processes is itself a draw from a parent Dirichlet process. HDP provides a nature way of modelling groups of data where the same atoms (clusters) are shared among all the groups. There are various applications of HDP, includes hierarchical topic model [43, 44, 45], human intracranial electroencephalogram [46], universal binary model [47], etc. Nesting is another way to supply prior for multilevel cluster structure. One attractive works is nested Dirichlet process (nDP)[48] which replaces the random atoms of DP with random probability measures drawn from a DP. In other words, the base measure that nDP uses is itself a DP. NDPs have several extensions that combine nDPs with HDPs, such as nested CRFP(nCRFP) [49], nested hierarchical Dirichlet Processes [50], hybrid nested/ hierarchical Dirichlet process (hNHDP) [51].

Both hierarchy and nesting supply good tools to capture the multi-layer structure. Nevertheless, their difference is significant. The distributions generated by HDP share the same atoms but with different weights while the different distri-

butions drawn from nDP have either the same atoms with the same weights or completely different atoms and different weights.

One commonly used approach for introducing non-exchangeability is to use appropriate distribution/processes to capture dependency in the atoms/weights in **stick-breaking representation** of DP and HDP. One of the pioneering works is order-dependent DP (πDDP) [52], which uses a point process to control the assignment of the weights and atoms. Transformed DP (tDP)[26] extends HDP by introducing random transformations for the global atoms which are shared in the sub-groups. Dynamic HDP [53] adds extra innovation distributions and random parameters to HDP to control the time-evolving weights for shared atoms across groups. Another interesting work is Hierarchical Dirichlet scaling process [45] that scales the mixture components using gamma representation of HDP.

Another approach to model the non-exchangeability in clustering is to capture the correlation and influence between observations. Such kind of works usually employ **Chinese restaurant process(CRP)** representations to define the dependency between observations. A representative model is distance dependent Chinese restaurant process (dd-CRP)[9], which we have reviewed in detail. Similarly, spectral CRP [54] uses the similarity between documents to map them into a low-dimensional spectral space where we then compare several clustering methods. Region-based dd-CRP(rdd-CRP) [55] generalizes dd-CRP to a hierarchical structure, sharing the cluster components across groups while allows within-group clustering to depend on external distance measurements. Another important work of this type is recurrent Chinese restaurant process (rCRP) [56] that introduces epochs that evolve over time in a Markovian fashion to capture temporal order of observation for evolutionary clustering.

2.3 Beta Process and Its Variants

Beta process was originally developed for survival analysis on life history data [24]. It was utilized as a prior distribution over the space of cumulative hazard function. As its name suggests, it produces cumulative hazard rates whose increments are independent and beta distributed. Later, the work in [57] extended beta process to more general spaces for different applications, such as factor analysis [58], dictionary learning [59, 60], image interpolation and document analysis [57].

2.3.1 Beta Process

On a measurable space Ω, a beta process H is defined as a positive Levy process, a positive random measure whose masses on disjoint subsets of Ω are independent. It is parameterised by a positive concentration function c and a base measure H_0, which is also defined on space Ω. In simplified cases, where function $c(\omega_i)$ becomes a constant, we call c concentration parameter.

For disjoint infinitesimal partitions of Ω, the beta process can be generated as:

$$H(B_k) \backsim Beta(cH_0(B_k), c(1 - H_0(B_k)), \tag{2.14}$$

where B_k indicates a partition, and $k \in \{1, \cdots, K\}$ is the index. The process can be denoted as $H \backsim BP(c, H_0)$.

When the base measure H_0 is discrete and has a set function form of $H_0 = \sum_i p_i \delta_{\omega_i}$, H turns to have atoms at the same locations as H_0's and can be written in a set function form accordingly as:

$$\begin{aligned} H(\omega) &= \sum_i \pi_i \delta_{\omega_i}(\omega) \\ \pi_i &\backsim Beta(cq_i, c(1 - q_i)), \end{aligned} \tag{2.15}$$

where $\delta_{\omega_i}(\omega) = 1$ when $\omega = \omega_i$ and 0 otherwise.

It is worth noting that π_i does not serve as a probability mass function on Ω. It is a part of beta process, a positive random measure over Ω. It also helps beta process to parameterize Bernoulli process (see Sec. 2.3.1.1).

With beta process as a non-parametric prior, the series of latent Bernoulli variables become a Bernoulli process which is useful to construct infinite latent model.

2.3.1.1 Bernoulli Process

For a Bernoulli process $BeP(H)$, each of its draws X_j is again a measure on space Ω. j represents the draw index. H indicates a beta process on Ω, as defined before. It acts as the prior of the Bernoulli process. A draw of the Bernoulli process can also be represented via a set function form as:

$$\begin{aligned} X_j(\omega) &= \sum_i x_{ij}\delta_{\omega_i}(\omega) \\ x_{i,j} &\sim Bernoulli(\pi_i), \end{aligned} \tag{2.16}$$

where δ_{ω_i} corresponds to the same atom location of H. The random variable x_{ij} is generated from a Bernoulli distribution parameterized by π_i which is defined as Eq. 2.15. With x_{ij} as its elements, an infinite binary column vector, also denoted by X_j, can be used for representing a draw of the Bernoulli process. Then the draws of the Bernoulli process can form an infinite binary matrix X, with X_j representing a column and j representing the column index. Each row of the matrix $X_{:,i}$ corresponds to an atom location δ_{ω_i}.

$$\begin{aligned} H &\sim BP(c, H_0) \\ X_{:,i} &\sim BeP(H). \end{aligned} \tag{2.17}$$

It is worth noting that beta process is a conjugate prior of Bernoulli process. Given a beta process prior and a set of m observations drawn from a Bernoulli process,

the posterior is again a beta process, with parameters updated as follow:

$$H|X_{1,\cdots,m} \backsim BP(c+m, \frac{c}{c+m}H_0 + \frac{1}{c+m}\sum_{j=1}^{m} X_j). \tag{2.18}$$

The conjugacy significantly simplifies the inference procedure for parameter estimation. We can see that the beta process appears to be a proper Bayesian non-parametric prior for such infinite binary matrices.

2.3.2 Hierarchical Beta Process

Hierarchical beta process(HBP) was originally proposed to model the complex hierarchical structure in documents classification[57]. In a manner akin to Hierarchical Dirichlet process, a hierarchy over beta process is proposed

$$\begin{aligned} B &\sim BP(c_0, B_0) \\ B_j &\sim BP(c_j, B) \\ X^j_{:,i} &\sim BeP(B_j), \end{aligned} \tag{2.19}$$

where $X^j_{:,i}$ denotes the i^{th} observation in j^{th} group. Similar to hierarchical Dirichlet process that allows group of data to share the atoms across group, HBP has proved a good choice as a non-parametric prior in latent factor analysis for multiple groups of data [61], where the factorial subspace is shared across the group.

Because the property of Levy process, the space ω can be partitioned into discrete part and continues part, and inference can be performed separately on each part[57]. For the discrete part, we have:

$$\begin{aligned} b &\sim Beta(c_0, b_0) \\ b_j &\sim Beta(c_j, b) \\ x_{ij} &\sim Ber(b_j), \end{aligned} \tag{2.20}$$

where $b_0 = B_0(\omega)$, $b = B(\omega)$, $b_j = B_j(\omega)$, $x_{ij} = X_{i,j}(\omega)$, Let $m_j = \sum_{i=1}^{n_j} x_{ij}$. The log posterior distribution of b given x can be obtained by marginalization out b_j:

$$\begin{aligned} f(b|x) = (c_0 b_0 - 1)log(b) + (c_0(1 - b_0) - 1)log(1 - b)) \\ + \sum_{j=1}^{N}(\sum_{i=0}^{m_j - 1} log(c_j b + i) + \sum_{i=0}^{n_j - m_j - 1} log(c_j(1 - b) + i)). \end{aligned} \tag{2.21}$$

The posterior is log concave and thus can be maximized at $b \in (0, 1)$.

2.3.3 Other Variants of BP

Beta process and its construction representation, Indian buffet process (IBP)[62], have become models of choice in building non-parametric Bayesian latent feature model where the observations are represented using an unknown number of latent features.

In non-parametric latent feature learning model, BP or IPB is adopted as a prior of the binary matrix that governs the feature allocation with an infinite number of exchangeable columns. The linear-Gaussian formulation or its extension is used as the likelihood function, which relates the binary feature allocation matrix to observation with assistance of a feature value matrix that governs the scale of the features. This model is applied in various domains, including dyadic data modelling [63], non-parametric factor analysis[58], non-parametric Bayesian dictionary learning for sparse image [59], non-parametric link prediction[64] and infinite canonical correlation Analysis [65], etc.

A distinctive application of BP is infinite hidden causes model [66], where BP is used as the latent cause allocation binary matrix while the combination impacts of causes on the observed variables is modelled via a Noisy-Or [67] distribution.

Akin to DP, beta process assumed that the observations are infinitely exchangeable which is not valid in many cases. Therefore, various extensions of BP are

proposed to solve this issue. Kernel beta process [68] introduces the covariate dependence through a kernel that defines the distance of feature and sample in covariate space. Dependent Hierarchical Beta Process [60] removes the local exchangeability in HBP by introducing a covariates kernel that captures the relationships between samples.

Other works that introduce non-exchangeability are based on IBP representation. The common approach is to use a distribution to control the generation process of feature allocation matrix entries. Phylogenetic Indian buffet process(pIBP) [69] summarizes the prior knowledge about the dependency of observations using a tree structure to control the activation of the matrix entries. Distant dependent Indian buffet process[70] biases nearby data to share more features by using the connection between data points based on distance. Dependent Indian buffet process [71] capture the latent covariate dependence of data var a hierarchical Gaussian process.

2.4 Interaction Point Processes

Stochastic point process [72] is a collection of mathematical random points falling in some underlying mathematical space. It is designed for tackling various temporal event modelling problems. In most applications, the point represents the time and/or location of an event. Point process is useful in modelling a wide variety fields of natural phenomena, including epidemiology, finance, ecology, forestry, mining and meteorology.

Stochastic Point Process: If a sequence of random variables $T = t_1, t_2, ..., t_n$, taking values in $[0, \infty)$, has $P(0 \leq t_1 \leq t_2 \leq ...t_n) = 1$, and the count of points $N(t)$ in a bounded region is finite, then T is a (simple) point process. Correspondingly, $N(t)$ is called counting process.

A stochastic point process can be defined via its conditional intensity function that provides an equivalent representation as a counting process for temporal events.

Consider a point process $T = \{t_i\}_{i=1}^{N}$ with the count $N(t)$ and the associated histories $H(t)$. If a (non-negative) function $\lambda^*(t)$ exists, which has such form

$$\lambda^*(t) = \lim_{\Delta t \to 0} \frac{E[N(t + \Delta t) - N(t)|H(t)]}{\Delta t}. \tag{2.22}$$

The conditional intensity function $\lambda^*(t)$ of point process T can be intuitively interpreted as the average number of points arrived per unit time at t. Clearly, $\lambda^*(t)$ only depends on the history information $H(t)$.

Homogeneous Poisson Process is the simplest and most ubiquitous point process model. Its conditional intensity is a constant.

$$\lambda^*(t) = \lambda \tag{2.23}$$

There is no interaction between events because the current intensity is not affected by historic occurrence of points. Thus, once the number of event is determined, the locations of points are independently distributed. Homogeneous Poisson process has been a basis of the investigation of many point process. However, the expressiveness of it is fairly limited as many events in our world are non-independent.

Interaction point process (IPP) [1] [73, 74] defines a broad range of stochastic point processes that can model various interaction mechanisms. In IPP, the interaction between two points is described by a pair interaction function, usually a function of the inter-points distance in the observation space. An IPP is called *self-exciting* if each event's arrival increases the rate of future event arrivals and *self-correcting* if on the contrary decrease the future arrivals.

2.4.1 Hawkes Process

Hawkes process is one of the most general and flexible *self-exciting* IPPs, named after its inventor Alan G. Hawkes [75, 76]. In Hawkes process, an event tends to

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

'trigger' the future events, by which it means each arrival increases the probability of subsequent arrivals for a period of time. Hawkes process can be defined in two equivalent ways: definition using conditional intensity function and definition as Poisson cluster process.

2.4.1.1 Definition by Conditional Intensity

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:

$$\lambda^*(t) = u(t) + \sum_{i:t_i<t} \alpha\beta(t - t_i), \tag{2.24}$$

where, $\mu(t)$ is a non-negative function on R, called background intensity. The product $\alpha\beta(t - t_i)$ represents the overall excitation intensity, in which, parameter α describes the degree that each historic arrival influences the intensity while function $\beta(t)$ which is defined on $[0, \infty)$ indicates how this influence lasts by time [2].

Typical $\beta(t)$ selections are in decay function forms, *e.g.*, exponential decay function[75] and power law decay function[77] etc. Thus, the triggering effect of a point appears immediately after its occurrence and quickly decays in certain ways, thereby showing clustering patterns. It is worth noting that we use $\lambda(t)$ to represent intensity function conditioned on previous points with the consideration of notation simplicity.

Assume we have observed a set of data points $X = \{t_i\}_{i=1}^{N}$ that distributed on Hawkes process. The likelihood function [72] is given by :

$$L(X|u, \alpha, \beta) = \left(\prod_{i=1}^{n} \lambda^*(t_i) \right) exp\left(-\Lambda(T)\right), \tag{2.25}$$

[2]It is worth noting that we do not consider edge effect in this thesis, hence no events exist before time 0.

where,

$$\Lambda(T) = \int_0^t \lambda^*(s)ds = \int_0^t u(s)ds + \sum_{i:t_i<t} \alpha B(t - t_i), \tag{2.26}$$

where, $B()$ is the cumulate distribution function of $\beta()$.

If the prior of parameter is denoted by $p(u, \alpha, \beta)$, the posterior is derived by

$$p(u, \alpha, \beta|X) \propto p(u, \alpha, \beta)L(X|u, \alpha, \beta). \tag{2.27}$$

2.4.1.2 Poisson Cluster Process

Because the superposition of several Poisson processes is also a Poisson process, Hawkes process can also be viewed equally as a Poisson cluster process that is constituted by a background Poisson process and a collection of triggered Poisson processes following a certain branching structure [76, 78].

The branching structure can be intuitively described by a metaphor of counting the population in a country where people are either immigrated from abroad or by birth. The arrival of immigrants to the country follows a homogeneous Poisson process at intensity $u(t)$ and each individual immigrant independently breeds zero or more number of children that is modelled by Poisson process. Fig. 2.2 illustrates this immigration-birth description.

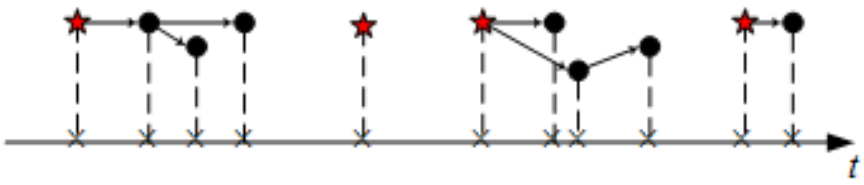

Figure 2.2: Immigration birth representation

In this representation, a Poisson cluster process consists of two types of points: immigrants (denoted by red pentacles in Fig. 2.2) and offsprings (denoted by bank dots in Fig. 2.2). 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 organised 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 $C = \{c_j\}$, where $c_j = i$ 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, Poisson cluster process is recursively defined. Consequently, more than one generation of descendants can be generated.

Based immigration birth representation, the likelihood of function of Hawkes process can be written with a multiplying of two parts: immigrant term and offspring term. given by:

$$L(X|y,u,\alpha,\beta) = p(I|y,u,\alpha,\beta)\prod_{i=1}^{n} p(O_i|u,\alpha,\beta), \tag{2.28}$$

where, y represents the branching structure. The immigrant term and the offspring term are respectively given by:

$$\begin{aligned} p(I|y,u,\alpha,\beta) &= \left(\prod_{t_i\in I}^{n} u(t_i)\right) exp\left(-\int_0^t u(s)ds\right) \\ p(O|y,u,\alpha,\beta) &= \left(\prod_{t_j\in O}^{n} \alpha\beta(t_j - t_{p(j)})\right) exp\left(-\sum_{t_i\in I} \alpha B(T-t_i)\right), \end{aligned} \tag{2.29}$$

where, $t_{p(j)}$ denotes the parent of t_j

2.4.2 Other Interaction Point Processes

While Hawkes process addresses an important case in which an occurrence of a point can cause additional points in near future, there exist other types of IPPs. *e.g.*, Hawkes process [75, 76], cascades of Poisson processes [14] and Neyman-Scott process [79]. The work in [80] provides a brief summary of some popular IPPs.

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 [14] 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 [79] is a Poisson cluster process that only allows one generation of offspring. For repulsive point processes [80, 81], which show inhabitation behaviours, intensity increment is suppressed once a point 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 [14], periodic activity can also be modelled 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 [75, 82]. Hence, all these IPPs can be unified and generalized with the support of the Poisson cluster process.

Poisson cluster process representation of IPPs is an important characteristic, which can be adopted to derive more expressive extensions of point process model by placing different prior on cluster structure. It institutes the basis of the contribution in Chapter 4 and Chapter 5.

2.5 Hidden Markov Model

The HMM [83, 84] is one of the most popular approaches for temporal event modelling. It has be applied in various fields where the underlying state sequence is not observable, such as speech recognition [85, 83], natural language processing [86], finance[87, 88, 89], etc.

HMM utilizes a sequence of latent states with Markovian property to model the dynamics of temporal events. Each event in the HMM is associated with a latent state that determines the event's observation via an emission probability distribution. The state of an event is independent of all but its most recent predecessor's state (*i.e.*, Markovianity) following a transition probability distribution.

The HMM consists of four components: (1) an initial state probability distribution,(2) a finite latent state space, (3) a state transition matrix and (4) an emission probability distribution. As a result, the inference for the HMM involves four corresponding parts. The graphic model is shown in Fig. 2.3.

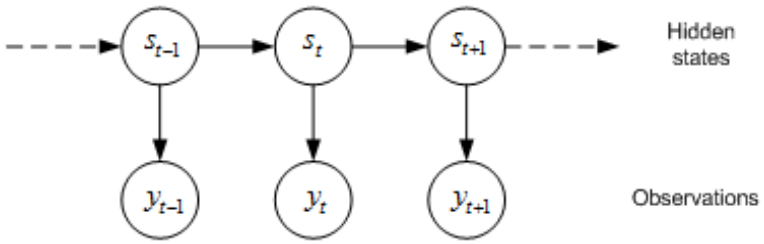

Figure 2.3: Hidden Markov model

The joint distribution of a sequence of latent states and observations can be factorized as:

$$p(S, Y|\Theta) = p(s_1|\pi)p(y_1|s_1)\prod_{t=2}^{N} p(s_t|s_{t-1}, A)p(y_t|s_t, \phi), \tag{2.30}$$

where, A denotes the sate transition matrix, ϕ denotes emission distributions parameter, π represents the initial latent state distribution parameter. Y is the observations and S is the latent states.

HMM can be viewed as a typical dynamic variant of a finite mixture model[10], in which the mixture component corresponds to the emission models and the latent variable corresponds to the hidden state. So it is quite natural to consider setting a Dirichlet process related prior on the mixture underlying the HMM, yielding a nonparametric model that can automatically infer the states' number based on observations.

2.6 Markov Chain Monte Carlo

The key task in exact Bayesian inference is to evaluate the posterior distribution $p(\Theta|X)$ of the parameters Θ given the observation X, such as estimating its expectation. In many cases, with a conjugation prior, the closed form of posterior distribution can be obtained by marginalization of the uninterested variables. However, in most circumstance, it is infeasible to marginalize variables as the lacks of conjugacy, thus the closed form of posterior distribution cannot be obtained. For example, for many Bayesian model has high dimension, it is possible to construct the form of equation for the posterior $p(\theta|x) \propto \tilde{p}(\theta, x)$, but infeasible to normalize the equation duo to the intractability of $\int \tilde{p}(\theta, x)dx$.

Markov chain Monte Carlo (MCMC) methods provides an effective tool to estimate the required distribution. It uses a Markov chain mechanism to generate samples to explore the objective space. The samples are drawn in a manner that emulates the sampling from the real target distribution. One of the significant advantages of MCMC lies on the fact that it does not rely on the form of the distribution. Actually, it can be with arbitrary complex form.

2.6.1 The Metropolis-Hastings Algorithm

The Metropolis-Hastings(MH) algorithm is one of the most popular examples of MCMC method. Most of the practical MCMC algorithms can be interpreted as

special cases or extensions of it.

The basic problem that MH solves is to sample from a distribution $p(x)$ from a function $\tilde{p}(x)$ that is proportional to the density of $p(x)$. Suppose we have current state of the Markov chain x_n, and we want to sample the next state x_{n+1}. The drawing can be described into two stages. The first stage is to generate a candidate x^* drown from a proposal distribution $q(x^*|x_n)$, which depends on the current state.

The second stage is to accept or reject the candidate. If accepted, the candidate will be the next state, otherwise use the current state value as the next state.

$$x_{n+1} = \begin{cases} x^* & accept \\ x_n & reject, \end{cases} \tag{2.31}$$

where the acceptance probability $A(x_{n+1} \longrightarrow x^*)$ is:

$$A(x_{n+1} \longrightarrow x^*) = \min\left(1, \frac{\tilde{p}(x^*)}{\tilde{p}(x_n)} \frac{q(x_n|x^*)}{q(x^*|x_n}\right). \tag{2.32}$$

Note $\frac{\tilde{p}(x^*)}{\tilde{p}(x_n)}$ does not depend on the normalizing constant $X_p = \int \tilde{p}(x)dx$. Thus $\tilde{p}(x)$ can be any un-normalized distribution over of x. The ratio $r_A = \frac{\tilde{p}(x^*)}{\tilde{p}(x_n)} \frac{q(x_n|x^*)}{q(x^*|x_n}$ is often called acceptance ratio.

This procedure essentially defines a Markov chain on x whose stationary distribution is $p(x)$. Although theoretically the proposal distribution $q(x)$ can be with any form as long as it is feasible to draw candidates x^*, the selection of proposal distribution has a significant impact on the algorithm performance. For a continuous states space, a typical option is to use a normal distribution centred on the current state x_n.

2.6.2 Gibbs Sampling Algorithm

Gibbs sampling [90] is a simple and widely applicable MCMC algorithm and can be seen as a special case of the Metropolis Hastings algorithm. It draws random samples based on the conditional distributions of variables.

Consider we wish to sample distribution from $p(\mathbf{z}) = p(z_1, ..., z_i, ..., z_M)$. Firstly, some values for the variables are set as the initial state of the Markov chain. Then the variable vector $\mathbf{z}$ is updated iteratively. In each iterative, the algorithm updates variable z_i with a new draw from a conditional distribution on the other variables.

$$z_i \sim p(z_i|\mathbf{z}_{\imath}) \qquad for \quad i \in [1, M], \tag{2.33}$$

where $\mathbf{z}_{\imath}$ represent $z_1, ..., z_M$ but with z_i omitted.

Compared with MH algorithm, which "corrects" the Markov chain by "accepting" or "rejecting" strategy to ensure the invariant distribution can converge to the target distribution $p(x)$, Gibbs sampling is simpler while providing a general framework that simplifies high dimensional sampling by successively sampling each dimension based on the conditional distribution. In Bayesian inference, this characteristic is greatly favourable as Bayesian models can be typically defined via a collection of conditional distributions.

Chapter 3

Bayesian Nonparametric Approach for Sparse Event Prediction

In this chapter, I propose a Bayesian nonparametric approach, namely the Dirichlet process mixture of hierarchical beta process, to handle the sparsity problem occurred in many real-world temporal event modelling scenarios. In these scenarios, the occurrence of temporal events are rare, which causes the lack of event samples for model training. Particularly, in the area of infrastructure asset (e.g., water pipe network) predictive maintenance, we aim to accurately predict future asset failures so that preventative maintenance can be wisely planned and conducted to avoid disastrous failures. However, the failures of asset are usually rare. Many assets even have no failure events in the history for training. Hence, a modelling method that can handle sparse events becomes critical for these scenarios. With the support of BNP, the chapter demonstrates that the sparsity problem can be well tackled via the mechanism of flexible clustering and event sharing.

In this chapter, I use water pipe failure prediction as an concrete application example to present the proposed method for solving the sparsity problem. How-

ever, the model itself is general enough to be applied to other spare temporal event modeling problems.

3.1 Introduction

Water supply networks are valuable urban infrastructure assets that are responsible for reliable water resource distributions. However, as urbanization trends continue and urban populations rise, water utilities find it increasingly difficult to meet growing water demand with aging and failing water pipe networks. Water pipe failures, which can cause tremendous economic and social costs, as illustrated in Fig. 3.1, have become the primary challenge to water utilities. In order to tackle the problem in a financially viable way, preventative risk management strategies are widely adopted by water utilities to prevent disastrous failures. The basic idea of the strategies is to identify high-risk pipes proactively and renew them in time to avoid potential failures. Meanwhile, it is also required to avert any replacements of pipes that are still in healthy condition. Therefore, the strategies consist of two main steps accordingly: (1) high-risk pipe prioritization, in which pipes are ranked based on their risks of failures, and (2) physical condition assessment, in which physical inspections are conducted on highly rated pipes to confirm their actual conditions for replacements. The pipes, which are not identified as high-risk pipes at the prioritization step, will only be renewed reactively. Hence, the success of the strategies heavily relies on the prioritization step. For making accurate selections of high-risk pipes, the prioritization step requires a failure prediction method that can give a precise estimation of pipe failure likelihood, based on which the estimated failure cost and renewal cost can be readily obtained.

The mechanisms of water pipe failures have been studied for many decades, and two main research avenues exist for water pipe failure prediction, namely physical modelling and statistical modelling. For physical modelling, a variety of models has been proposed for explaining and predicting the deterioration processes of water

Figure 3.1: Scenes of water pipe failures.

pipes, such as pipe-soil interaction analysis, residual structural resistance, corrosion status index and hydraulic characteristics modelling. A comprehensive review can be found in [91]. For statistical modelling, it considers historical failure records, pipe attributes and environmental factors together for making predictions. It assumes that pipes with similar intrinsic attributes and environmental factors share similar failure patterns, and that failure patterns which appeared before are likely to reappear in the future.

Although physical models can help understand the mechanisms of water pipe failures, they have strong limitations when being applied to large-scale pipe networks for failure prediction[1]. In order to improve high-risk pipe prioritization for large-scale metropolitan pipe networks, I propose a Bayesian nonparametric statistical approach, namely the Dirichlet process mixture of hierarchical beta process model, for water pipe failure prediction. Unlike parametric approaches, the structure and complexity of the proposed model can grow as the amount of observed data increases. It makes the model invulnerable to faulty assumptions of model forms and adaptable to various failure patterns, thereby leading to more accurate predictions for different application scenarios.

It is worth noting that water pipe failure data is extremely sparse in reality. Very few pipes have failure records during the observation period. Such sparsity makes traditional failure prediction methods incompetent for accurate pipe failure

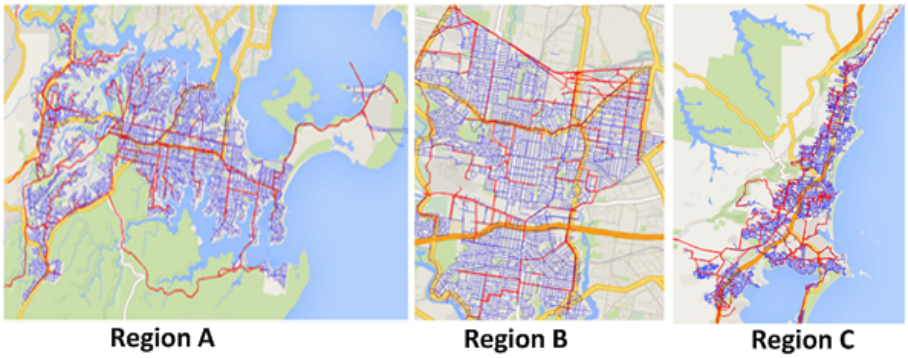

Figure 3.2: Water supply networks in the selected regions.

prediction since most pipes do not have failure data for training. The proposed approach deals with this issue by sharing failure data via a flexible hierarchical modelling of failure behaviours. The key component of the hierarchical modelling is a flexible grouping scheme. It clusters similar pipes together for modelling so that failure data can be shared by similar pipes for training. Additionally, the failure data sparsity is exploited for developing an approximated yet computationally efficient Metropolis-within-Gibbs sampling method for model parameter inference.

The proposed method has been applied to the water supply network of an international metropolis that has a total population of near five million people. In this work, three representative regions are selected from the metropolis for comparison experiments. The regions and the networks are shown in Fig. 3.2. As I can see, the water supply network is constituted of two main categories of water pipes, critical water main (CWM) indicated by red lines and reticulation water main (RWM) indicated by blue lines. CWMs have larger diameters (300 millimetres and above), and RWMs have smaller diameters (smaller than 300 millimetres). Each water pipe is composed of a set of pipe segments connected in series. Failure records can be precisely matched with pipe segments, allowing the proposed method to model failure behaviours of pipe segments.

Although CWM only takes a small portion of the network, it plays a more important role than RWM. The consequences caused by CWM failures are significantly

more severe than RWM failures. Given the large scale of the network and the high cost of physical inspection, it is infeasible to inspect both types of pipes physically. Thus, the preventative risk management strategy just focuses on preventing CWM failures in practice. Accordingly, in this work, the proposed approach is compared with other state-of-the-art prediction methods for predicting CWM failures, of which the data sparsity is even more significant. The experimental results obtained from the real-world pipe data demonstrate the superiority of the proposed approach over the others.

The rest of the paper is organised as follows. Section 3.2 reviews the related work. Section 3.3 describes the details of the proposed method. Empirical studies are shown in Section 3.5, and the conclusions are drawn in Section 3.6.

3.2 Statistical Failure Prediction Methods

In recent decades, many statistical models have been proposed for water pipe failure prediction. In the early stages, various methods were developed for modelling the relationship between pipe age and pipe failure rate. For instance, the work in [92] proposed a time-exponential model, which formulates the number of failures per unit length per year as an exponential function of pipe age. Similarly, time-power model [93] and time-linear model [94] were developed with comparable performances.

Later on, multivariate probabilistic models were suggested. They make predictions based on a variety of pipe attributes, such as age, material, length and diameter. One of the most popular multivariate approaches is the Cox proportional hazards model [7]. It is a semi-parametric method, in which the baseline hazard function has an arbitrary form and the pipe attributes alter the baseline hazard function via an exponential function multiplicatively. The Weibull model and its variants [95, 6] are also widely adopted in practice. They utilize either a Weibull distribution or a Weibull process for modelling pipe failure behaviours.

Recently, a ranking-based method [96] was proposed for predicting water pipe failures. It treats failure prediction as a ranking problem. Pipes are ranked based on their failure risk. The method performs failure prediction via a real-valued ranking function rather than an estimation of failure probability.

3.3 The Proposed Method

The proposed Dirichlet process mixture of hierarchical beta process model consists of two main components working with each other interactively: a hierarchical representation of water pipe failure behaviours and a flexible pipe grouping scheme. The grouping scheme generates a set of groups, on each of which the hierarchical representation can be constructed. The hierarchical representation provides a precise modelling of each group's failure behaviours, hence acts as the basis of grouping.

The two main components are described in Section 3.3.1 and Section 3.3.2 respectively. The details of the proposed model are given in Section 3.3.3. The inference algorithm is developed in Section 3.4.

3.3.1 Hierarchical modelling of Water Pipe Failure Behaviours

The hierarchical beta process is adopted in this work as the hierarchical prior of water pipe failure events which themselves can be modelled by Bernoulli process. The details of the hierarchical failure model are illustrated in Fig. 3.3.

With the aid of beta-Bernoulli process, a hierarchical representation can be developed for modelling water pipe failure behaviours. Firstly, failure events can be modelled by a Bernoulli process $BeP(H)$. Let an infinite binary matrix X, as illustrated in Fig. 3.3, represent failure records of pipes. Each of its columns, X_j, can be treated as a draw from the Bernoulli process $BeP(H)$. It is an infinite binary column vector with the i-th element $x_{i,j}$ generated from $x_{i,j} \backsim Bernoulli(\pi_i)$.

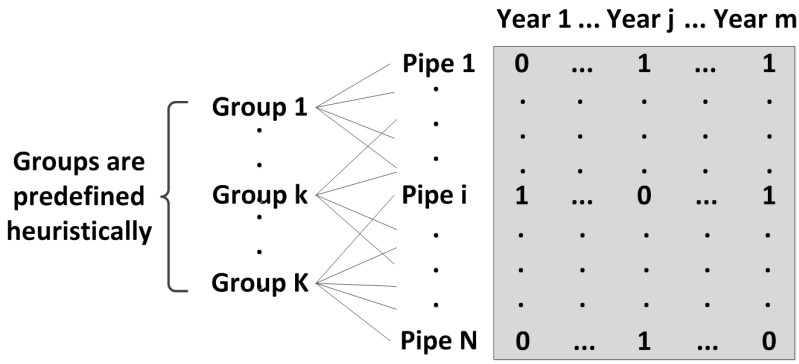

Figure 3.3: Binary failure matrices for pipes and pipe segments

$x_{i,j} = 1$ means pipe i failed in year j, and $x_{i,j} = 0$ otherwise. Then the beta process, $H \backsim BP(c, H_0)$, defined as a positive Levy process on pipe space Ω, can be used as a prior of failure events, namely failure probability. Its set function is defined as follows.

$$\begin{aligned} X_j(\omega) &= \sum_i x_{ij}\delta_{\omega_i}(\omega) \\ x_{i,j} &\sim Bernoulli(\pi_i). \end{aligned} \tag{3.1}$$

With beta process H as a prior, each row of the matrix X corresponds to an atom location δ_{ω_i} in the pipe space Ω, which can be infinitely large. We assume that two pipes share the same failure patterns if they have the same intrinsic attributes and environmental factors. Hence, we treat such two pipes as the same in the pipe space Ω. Considering all the possible combinations of pipe attributes and environmental factors, the number of "unique" pipes in the pipe space becomes infinite. Therefore, each column of the matrix X is an infinite binary vector that is drawn from a Bernoulli process. The beta process H is then a conjugate prior of the infinite binary matrix X. It models the failure probabilities of pipes via π_i.

While beta-Bernoulli process is capable of modelling failure behaviours as described above, there are two issues of adopting it in practice. Firstly, the number of failures is extremely small compared with the number of pipes, especially for CWMs. Only a small portion of CWMs have failure records since most of CWMs did not fail

during the observation period. Thus, the majority of CWMs have no failure data for model training. Secondly, in addition to pipe failure histories, pipe attributes and environmental factors are also crucial for estimating failure probabilities. However, they are not properly considered in beta-Bernoulli process. The fact that the pipes with similar intrinsic attributes and environmental factors often share similar failure patterns is ignored by beta-Bernoulli process.

In order to address these issues, the hierarchical beta process (HBP) model [8, 57] can be adopted as a hierarchical modelling of water pipe failure behaviours. Given a water pipe grouping, *e.g.*, grouping by intrinsic attributes, one more beta process can be added into the model hierarchy for modelling the failure behaviours of groups. The new beta process is on top of the existing beta process, serving as the prior of its mean parameter. The HBP model can also be described as the followings:

$$\begin{aligned} q_k &\backsim Beta(c_0 q_0, c_0(1-q_0)), \ k \in [1, \cdots, K], \\ \pi_i &\backsim Beta(c_k q_k, c_k(1-q_k)), \ i \in [1, \cdots, N], \\ x_{i,j} &\backsim Bernoulli(\pi_i), \qquad\qquad j \in [1, \cdots, m_i], \end{aligned} \tag{3.2}$$

where π_i and x_{ij} are defined as before, modelling the failure probability of pipe i and failure history of pipe i in year j respectively. q_k and c_k are the mean and concentration parameters for group k. q_k can be regarded as modelling the failure rate of group k. q_0 and c_0 are the hyper parameters.

By adding one more hierarchy level, the HBP model estimates failure probabilities through the inferences on both group level and pipe level. Group level inference estimates the group failure rate q_k, and pipe level inference estimates the pipe failure probability π_i. Failure data can be shared by the same group of pipes for estimating group failure rate q_k. It helps to solve the failure data sparsity problem. The failure patterns that are shared by similar pipes are captured at the group level since the pipes within the same group share the same q_k. At the pipe level, the pipe failure probability π_i is estimated by considering not only the failure observations x_{ij}, but

also the group similarity through the group failure rate q_k.

3.3.2 Flexible Water Pipe Grouping

Real world data is complicated and often demonstrates multi-modality property, which is the case for water pipe failures. Consequently, single-modality models become insufficient in such circumstances for modelling the whole data corpora. Mixture model is a widely adopted probabilistic approach for modelling the data arising from different modalities. It assumes that the final model consists of a set of mixture components, each of which can accurately model a portion of data.

For conventional parametric mixture models, the number of mixture components is required to be known in advance, which is unrealistic for many real world applications, such as water pipe grouping. Therefore, we adopt Dirichlet process (DP), a nonparametric approach, for pipe grouping. It serves as a flexible prior for data partitioning and sets no assumptions on the number of partitions. Correspondingly, the Dirichlet process mixture model, which is built based on Dirichlet process, can comprise a countably infinite number of components and adjust itself for fitting observed data.

In order to adopt DP as the prior of pipe grouping, we use the Chinese restaurant process (CRP) [97] as the constructive representation of DP. With the aid of the CRP, we can group pipes adaptively for fitting data observations. As a result, pipes with similar failure behaviours are grouped together. Moreover, the CRP helps to integrate the grouping process and the failure modelling process for achieving accurate performance.

3.3.3 Dirichlet Process Mixture of Hierarchical Beta Process

In this section, we give the detailed description of the proposed Dirichlet process mixture of hierarchical beta process (DPMHBP) model for water pipe failure prediction.

For the proposed DPMHBP model, a water pipe is treated as a set of pipe segments that are connected in series. The failure probability of a pipe segment is modelled by a beta process. It is different from the HBP model [8] where beta process is used for modelling failure probabilities of pipes.

Pipe length is an important attribute for estimating failure probability. The intuition is that longer pipes tend to have higher failure probabilities if other attributes and external factors are the same. However, the HBP model ignores the impact of the length attribute when estimating failure probabilities. It only focuses on pipe age attribute and failure histories. The significant variance of pipe lengths is neglected. In order to tackle the problem, the proposed approach suggests to model the failure probabilities of pipe segments whose lengths are relatively constant with a very small variance.

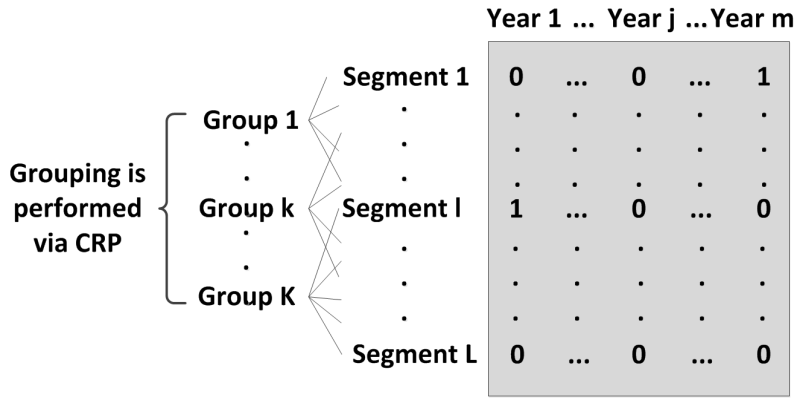

Figure 3.4: Binary failure matrices for pipes and pipe segments

Another difference between the HBP model and the proposed DPMHBP model is that the HBP model groups pipes based on heuristic domain information *e.g.*,

pipe age. Its grouping is predefined and fixed during the inference process. The number of the groups is also required to be set beforehand, which can be heuristic. In contrast, for the proposed DPMHBP method, the grouping process is integrated with the inference process via the DP mixture model. They interact with each other for achieving an optimal model. The number of groups is not fixed and can grow as the size of the training data increases.

Considering all the issues mentioned above, the DPMHBP model can finally be given as follows:

$$\begin{aligned} q_k &\backsim Beta(c_0 q_0, c_0(1-q_0)), && k \in [1, \cdots, K], \\ z_l &\backsim CRP(\alpha), && z_l \in [1, \cdots, K], \\ \rho_l &\backsim Beta(c_{z_l} q_{z_l}, c_{z_l}(1-q_{z_l})), && l \in [1, \cdots, L], \\ y_{l,j} &\backsim Bernoulli(\rho_l), && j \in [1, \cdots, m_l], \\ \pi_i &= 1 - \prod_{l=1}^{s_i}(1-\rho_l), && l \in [1, \cdots, s_i]. \end{aligned} \tag{3.3}$$

The failure probability estimation is conducted on three levels: segment group level, segment level and pipe level. The failure events are recorded for segments rather than pipes. The grouping is performed on segments via the CRP, as illustrated by Fig. 3.4. At segment group level, q_k denotes the failure rate of segment group k. z_l represents the group index for segment l. At segment level, ρ_l indicates the failure probability of segment l. Once the segment level estimation is obtained, pipe failure probability π_i can be readily computed via the failure probability of a series of connected segments. Fig. 3.5 shows the graphical model of the DPMHBP model.

It is worth noting that Bernoulli process is more suitable for modelling segment failures than modelling pipe failures because it is very rare for a segment to fail twice in a year.

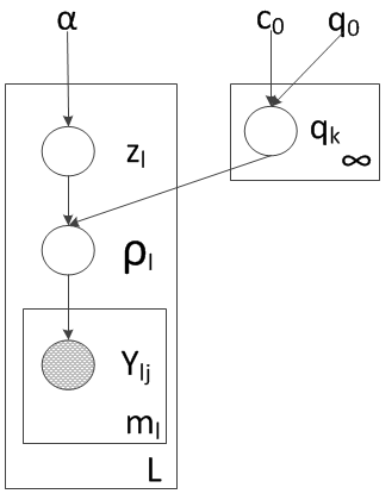

Figure 3.5: Graphical models for HPMHBP

3.4 Inference Algorithm

In this section, we describe the inference algorithm for the proposed DPMHBP model. Suppose we are given: (1) a set of pipes, $\{u_i\}_{i=1}^{N}$, (2) the segment composition $\{v_l\}_{l=1}^{s_i}$ for each pipe u_i, and (3) m-year failure records for segments, $\{y_{l,j}\}_{l\in[1,L],j\in[1,m]}$. The aim of the inference is to estimate pipe failure probabilities $\{\pi_i\}_{i=1}^{N}$, which are required for both physical condition assessment and proactive replacement. In order to achieve the goal, we need to estimate the variables $\{q_k\}_{k=1}^{K}$, $\{z_l\}_{l=1}^{L}$ and $\{\rho_l\}_{l=1}^{L}$.

Since no analytical solution is available for the proposed model, we use Markov chain Monte Carlo (MCMC) sampling algorithm for inference. Gibbs sampling is the MCMC-based method that has been widely used for DP mixture models when conjugacy exists between prior and likelihood. However, for the DPMHBP model, such conjugacy is broken by the extra hierarchy of the HBP model. Therefore, we choose to utilise Metropolis-within-Gibbs sampling method for inference.

For Metropolis-within-Gibbs sampling method, model variables are updated one by one iteratively until convergence as Gibbs sampling does. However, for each update step, Metropolis-Hastings sampling is used if the conditional distribution of

a model variable is not available for sampling. It is the case for $\{q_k\}_{k=1}^{K}$ and $\{z_l\}_{l=1}^{L}$.

To update each variables, because the conjugate prior for the likelihood function does not exist, Metropolis-Hastings sampling method is used to update group index z_l. The CRP conditional prior of z_l, defined by Eq. 2.9, is used as the proposal distribution for generating a candidate z_l^*. As a result, we can find that this factor cancels when computing the acceptance probability defined in Eq. 2.32 , leaving the acceptance ratio as:

$$r_{z_l} = \frac{p(y_{l,1\ldots m}|q_{z_l^*}, c_{z_l^*})}{p(y_{l,1\ldots m}|q_{z_l}, c_{z_l})}. \tag{3.4}$$

For the CRP conditional prior of z_l, defined in Eq. 2.9, k indicates the current number of segment groups, r denotes group index, α is the concentration parameter for CRP, n_r indicates the number of segments in group r, and z_{-l} represents all the z's with z_l removed. Non-informative prior is used for c_k's.

The likelihood function in r_{z_l} can be obtained by marginalizing out ρ_l:

$$\begin{aligned}
& p(y_{l,1\ldots m}|q_{z_l}, c_{z_l}) \\
= & \int p(y_{l,1\ldots m}|\rho_l) p(\rho_l|q_{z_l}, c_{z_l}) d\rho_l \\
= & \int p(\rho_l|q_{z_l}, c_{z_l}) \prod_{j=1}^{m} p(y_{l,j}|\rho_l) d\rho_l \\
= & \frac{\Gamma(c_{z_l})\Gamma(c_{z_l}q_{z_l} + \sum_{j=1}^{m} y_{l,j})}{\Gamma(c_{z_l}q_{z_l})\Gamma(c_{z_l}(1-q_{z_l}))} * \\
& \frac{\Gamma(c_{z_l}(1-q_{z_l}) + m - \sum_{j=1}^{m} y_{l,j})}{\Gamma(c_{z_l} + m)}.
\end{aligned} \tag{3.5}$$

Similar to z_l, Metropolis-Hastings sampling can be used for updating q_k. We use the normal distribution with the current value of q_k as mean for proposing a new candidate q_k^*. Hence, the Metropolis-Hastings sampling reduces to Metropolis sampling because of the symmetric proposal distribution. The acceptance ratio, r_{q_k}, can be

calculated accordingly as:

$$r_{q_k} = \frac{p(q_k^*)p(\{y_{l,1\ldots m}\}_{z_l=k}|q_k^*, c_k, \{z_l\}=k)}{p(q_k)p(\{y_{l,1\ldots m}\}_{z_l=k}|q_k, c_k, \{z_l\}=k)}. \tag{3.6}$$

The likelihood function in r_{q_k} can be computed by marginalizing out $\{\rho_l\}_{z_l=k}$ as:

$$\begin{aligned} & p(\{y_{l,1\ldots m}\}_{z_l=k}|q_k, c_k, \{z_l\}=k) \\ = & \prod_{l,z_l=k}\left[\int p(\rho_l|q_k, c_k, \{z_l\}=k)\prod_{j=1}^{m} p(y_{l,j}|\rho_l)d\rho_l\right] \\ = & \prod_{l,z_l=k}\left[\frac{\Gamma(c_k)\Gamma(c_k q_k + \sum_{j=1}^{m} y_{l,j})}{\Gamma(c_k q_k)\Gamma(c_k(1-q_k))}\right. \\ & \left.\cdot\frac{\Gamma(c_k(1-q_k) + m - \sum_{j=1}^{m} y_{l,j})}{\Gamma(c_k + m)}\right] \end{aligned} \tag{3.7}$$

Although the Metropolis sampling approach can sample new q_k, it is much less efficient than Gibbs sampling. Therefore, we derive an approximated Gibbs sampling step for updating q_k, in favour of the inference for large-scale datasets.

$$\begin{aligned} & p(q_k|c_k, \{z_l\}=k, \{y_{l,1\ldots m}\}_{z_l=k}) \\ \propto & \; p(q_k, \{y_{l,1\ldots m}\}_{z_l=k}|c_k, \{z_l\}=k) \\ = & \; p(q_k)p(\{y_{l,1\ldots m}\}_{z_l=k}|q_k, c_k, \{z_l\}=k) \\ \approx & \; \frac{\Gamma(c_0)}{\Gamma(c_0 q_0)\Gamma(c_0(1-q_0))} q_k^{c_0 q_0 - 1}(1-q_k)^{c_0(1-q_0)-1} \\ & \cdot \prod_{l,z_l=k}\left[\frac{(c_k q_k)^{s_l}\prod_{t=0}^{m-s_l-1}(c_k(1-q_k)+t)}{\prod_{t=0}^{m-1}(c_k+t)}.\right] \end{aligned} \tag{3.8}$$

The approximation made on the last step of Eq. 3.8 is based on the sparse nature of the pipe failure prediction problem. For water supply networks, most of the pipe segments never fail during the observation period. It is even more rarer for a pipe segment to have more than one failures during the observation period. Therefore, the number of failures for a segment, $s_l = \sum_j y_{l,j}$, satisfies $s_l \leq 1$, which leads to the approximation in Eq. 3.8.

By applying Taylor expansion, the posterior distribution in Eq. 3.8 can be further approximated as:

$$\begin{aligned} & p(q_k|c_k, \{z_l\} = k, \{y_{l,1\ldots m}\}_{z_l=k}) \\ \propto \; & q_k^{c_0 q_0 - 1}(1-q_k)^{c_0(1-q_0)-1} \\ & \cdot \prod_{l, z_l=k} \left[\left(\frac{c_k q_k}{c_k + m - 1} \right)^{s_l} (1-q_k)^{\sum_{t=0}^{m-s_l-1} \frac{c_k}{c_k+t}}. \right] \end{aligned} \tag{3.9}$$

Based on Eq. 3.9, we can see that the posterior of q_k can finally be approximated by a beta distribution:

$$\begin{aligned} & p(q_k|c_k, \{z_l\} = k, \{y_{l,1\ldots m}\}_{z_l=k}) \backsim \\ & \quad \text{Beta}\left(c_0 q_0 + \textstyle\sum_l s_l, c_0(1-q_0) + \sum_l \sum_t^{m-s_l-1} \frac{c_k}{c_k+t}. \right) \end{aligned} \tag{3.10}$$

It dramatically improves the efficiency of the updates for q_k.

For updating ρ_l, we can directly sample a new value from its conditional distribution given the other variables:

$$\begin{aligned} & p(\rho_l|q_{z_l}, z_l, c_{z_l}, y_{l,1\ldots m}) \backsim \\ & \quad \text{Beta}(c_{z_l} q_{z_l} + \sum_{j=1}^{m} y_{l,j}, c_{z_l}(1-q_{z_l}) + m - \sum_{j=1}^{m} y_{l,j}). \end{aligned} \tag{3.11}$$

Once, ρ_l is obtained, pipe failure probability π_i can be readily calculated via $\pi_i = 1 - \prod_l (1-\rho_l)$, $l \in [1, \cdots, s_i]$.

All the updating steps described above are performed iteratively until convergence is reached. Then the estimations for the model variables can be obtained by taking means of the sampled variable values, with burn-in samples omitted.

		# Pipes	# Failures	Laid years	Observation period
Region A	All	15189	4093	1930 – 1997	1998 – 2009
	CWM	3793	520	1930 – 1997	1998 – 2009
Region B	All	11836	3694	1888 – 1997	1998 – 2009
	CWM	2457	432	1888 – 1997	1998 – 2009
Region C	All	18001	4421	1913 – 1997	1998 – 2009
	CWM	5041	563	1913 – 1997	1998 – 2009

Table 3.1: Summary of pipe network data and pipe failure data .

3.5 Experiments

In this section, we conduct comparison experiments on the metropolitan water supply network data to demonstrate the superiority of the proposed DPMHBP model. We first introduce the pipe network data and the failure data in Section 3.5.1. The features used in the experiments are explained in Section 3.5.2. Then the compared methods are listed in Section 3.5.3. Finally, we give the comparison results and discuss the impact of the proposed method in Section 3.5.4.

3.5.1 Data Collection

Three representative regions from the metropolis are selected to perform the experiments. Region A is a local government area with a population around $210,000$, which is one of the most populous local government areas in its state. Its population density is 629 people per km^2. Region B is a local government area with a high population density of $2,374$ people per km^2. Its population is about $182,000$. Region C is a low density suburban local government area, which has a population of $205,000$ and a population density of 300 people per km^2.

For each region, both network data and failure data are collected. Network data consists of pipe IDs, pipe attributes, pipe locations and environmental factors. Pipe location is represented as a set of connected line segments, each of which corresponds to a pipe segment. Failure data contains pipe IDs, failure dates and failure locations.

Pipe amount, failure amount, laid year range and observation period are summarized for different pipe types in Table 3.1. As we can see, CWMs only take a small portion of the network, 24.97% for region A, 20.76% for region B, and 28.00% for region C. The ratio between CWM failures and all the failures is even more smaller, 12.71% for region A, 11.70% for region B, and 12.74% for region C.

The observation period covers 12 years, spanning from 1998 to 2009. It is short compared with pipe life span which can be more than 100 years as shown in Table 3.1. The majority of the pipes did not fail or just failed once during the observation period. If considering pipe segment failures, the failure events are even more sparser. Hence, the sparsity assumption holds for the proposed approximated sampling algorithm.

Failure locations are used for matching failures with pipe segments. It enables the proposed DPHBP model to work on pipe segment level for estimating failure probabilities.

As mentioned before, we focus on CWMs for comparison experiments since both physical condition assessment and proactive replacement are conducted for CWMs. For comparing the performances of different approaches, we use the first 11 years' failure records as training data and the last year's failure records as testing data. All the compared methods have the same setting for fair comparison.

3.5.2 Feature Description

In this section, we describe the pipe attributes and the environmental factors that we used in the experiments. There are five pipe attributes utilized in the experiments including protective coating, diameter, length, laid date, and material. Two types of environmental factors are considered in the experiments. One is the surrounding soil condition, and the other is the distance between pipe segment and its closest traffic intersection. These features are summarized in Table 3.2.

For pipe attributes, protective coating and material are categorical features indicating the type of coating and material. Typical protective coatings are polyethylene sleeve and tar coating. Typical materials are cast iron cement lined (CICL) and polyvinyl chloride (PVC). Diameter, length, and laid date are continuous features.

Surrounding soil condition is one of the most complex and important environmental factors for water pipe failure prediction. It puts direct impact on pipe degradation process. In the experiments, four different soil features are considered including soil corrosiveness, soil expansiveness, soil geology and soil map. They depict different perspectives of soil characteristics.

Soil corrosiveness describes the risk of pipe pitting (metal corrosion) which is essentially an electrical phenomenon and can be measured by linear polarization resistance test. Soil expansiveness describes the shrinking and swelling of expansive clays in response to moisture content change. It is a phenomenon that affects clay soil and can be measured by shrink swell test. Soil geology depicts the information of rocks, *e.g.*, sandstone and shale. Soil map represents the landscape information, *e.g.*, fluvial, colluvial and erosional. It also include the information of the soil types that are associated with different landscapes.

Each soil factor is a categorical feature containing several distinct values. The selected local government areas are partitioned into small regions according to the distinct values of soil factors. Pipe segments falling into the same region share the same soil factor value.

A large portion of CWMs are buried underneath roads. It makes the change of road surface pressure another important environmental factor for estimating water pipe failures. It has been shown that frequent pressure changes can lead to high failure rate. One of the main sources causing road surface pressure change comes from traffic intersections due to the frequent vehicle starting and stopping. In order to measure the impact of road surface pressure change, we calculate the distance between each pipe segment and its closest traffic intersection. The obtained con-

Property and factors	Description
Protective coating	Categorical value indicating the type of coating
Diameter	Continuous value indicating pipe diameter.
Length	Continuous value indicating pipe length
Laid date	Laid date for pipe
Material	Categorical value indicating the type of pipe material
Soil corrosiveness Soil expansiveness Soil geology Soil map	Categorical value indicating soil property for the corresponding soil factor
Distance to traffic intersection	Continuous value indicating the distance between pipe segment and the closest traffic intersection

Table 3.2: Pipe attributes and environmental factors

tinuous value is regarded as a feature of the pipe segment for predicting its failure probability.

3.5.3 Compared Approaches

In order to evaluate the proposed approach, four state-of-the-art methods are compared in the experiments including Cox proportional hazard model, Weibull model, HBP model and support vector machine (SVM) based ranking method. Additionally, different grouping methods are used with HBP model as comparisons for demonstrating the advantage of the grouping scheme of the proposed approach.

The Cox proportional hazard model [7] is one of the most popular approaches for survival analysis. It is a semi-parametric approach, in which the form of the baseline hazard function can be arbitrary, and the explanatory features put impacts on the baseline hazard function via an exponential function multiplicatively. Formally, the Cox proportional hazard model can be described as:

$$h(t, z) = h_0(t)e^{b^T z}, \tag{3.12}$$

where h_0 indicates the baseline hazard function, z indicates the explanatory features of water pipe, and b is the parameter vector that can be learned from training data via a partial likelihood maximization procedure.

For Weibull model [95, 6], water pipe failures are modelled as a set of stochastic events governed by a time dependent stochastic process, namely the Weibull process. It can be regarded as a nonhomogeneous Poisson process whose intensity varies as time changes. The intensity function can be formally given as:

$$\lambda(t) = \alpha\beta t^{\beta-1}, \tag{3.13}$$

where t represents pipe age, α and β are parameters that need to be learned from training data. Similar to Cox proportional hazard model, the explanatory features can also be utilized via an exponential function multiplicatively.

Analogous to the method proposed in [96], an SVM-based ranking approach is compared. This approach formulates pipe failure prediction as a ranking problem. It ranks pipes according to their failure risks without estimating their actual failure probability. It learns a real-valued ranking function H that maximizes the objective function:

$$\sum_{z\in P, z'\in N} \frac{I(H(z) > H(z'))}{|P| \cdot |N|}, \tag{3.14}$$

where P and N represent the positive class dataset (failure dataset) and negative class dataset respectively. $I(\cdot)$ is the indicator function. $|P|$ and $|N|$ indicate the numbers of data points in the positive and negative class datasets respectively.

The HBP model proposed by [8] is also compared. In order to evaluate the grouping scheme of the proposed approach, three different grouping methods are integrated with HBP model for comparisons. They group pipes based on pipe attributes according to domain expert suggestion. Specifically, pipes are grouped based on material, diameter and laid year.

For fair comparison, the features described in the previous section are used for

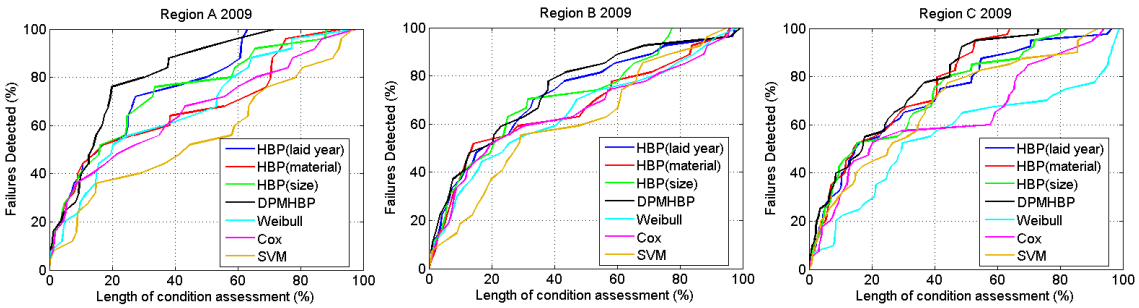

Figure 3.6: Failure prediction results for the selected regions by different models.

	Region A		Region B		Region C	
	DPMHBP	HBP	DPMHBP	HBP	DPMHBP	HBP
	Weibull	Cox	Weibull	Cox	Weibull	Cox
	SVM		SVM		SVM	
AUC(100%)	**82.67%**	77.05%	**74.51%**	72.56%	**78.37%**	73.54%
	68.44%	66.91%	65.20%	65.53%	55.84%	64.50%
	56.45%		61.90%		69.48%	
AUC(1%)	**8.09‰**	5.64‰	**4.21‰**	3.60‰	**5.11‰**	2.48‰
	5.84‰	4.67‰	2.70‰	2.46‰	2.98‰	2.50‰
	4.32‰		3.41‰		1.73‰	

Table 3.3: AUC of different approaches.

all the compared methods. For HBP and DPMHBP, the features are applied multiplicatively similar to Cox proportional hazard model and Weibull model. A linear kernel is used for the SVM-based ranking approach.

3.5.4 Prediction Results and Real Life Impact

In this section, we compare the prediction results to demonstrate the superiority of the proposed approach. As mentioned before, the historical failure data from 1998 to 2008 is used for training and the failures occurred in 2009 are used for testing. Water pipes are ranked by different methods based on their estimated failure risks. The failure prediction results are shown in Fig. 3.6. The x-axis represents the cumulative percentage of the inspected water pipes, and the y-axis indicates the percentage of the detected pipe failures.

	Region A		Region B		Region C	
	vs.HBP	vs.Weibull	vs.HBP	vs.Weibull	vs.HBP	vs.Weibull
	vs.Cox	vs.SVM	vs.Cox	vs.SVM	vs.Cox	vs.SVM
AUC(100%)	2.56	9.37	3.12	22.01	7.83	43.55
	(= 0.08)	(< 0.05)	(= 0.05)	(< 0.05)	(< 0.05)	(< 0.05)
	10.58	18.88	21.17	30.11	26.08	15.63
	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)
AUC(1%)	44.29	40.46	1.26	4.64	65.90	53.43
	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)
	62.44	69.01	5.53	1.99	65.43	61.72
	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)	(< 0.05)

Table 3.4: Statistical significance test (t-test) results.

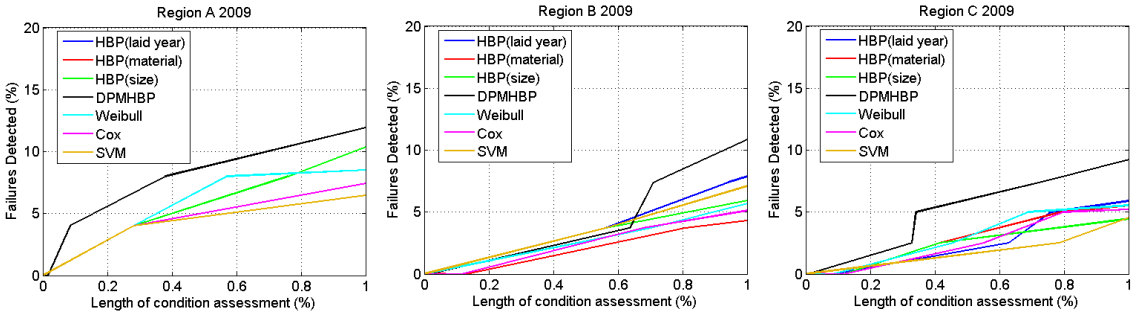

Figure 3.7: The detection results with 1% of pipe network length inspected.

Besides, we calculate AUC for measuring the performances of different approaches. The results are shown in Table 3.3. Statistical significance tests, particularly the one-sided paired t-test at 5% level of significance, are performed on AUC to evaluate the significance of the performance differences. The results are shown in Table 3.4. For Table 3.3 and Table 3.4, only the results from the best groupings are shown for the HBP model.

As we can see from Fig. 3.6 and Table 3.3, the proposed DPMHBP model consistently gives the most accurate prediction for all the three regions, whereas the other methods only perform accurately for some of the regions. It demonstrates the adaptability of the proposed approach to the diversity of failure patterns. The significance test results, listed in Table 3.4, show that the proposed model significantly outperforms the other methods.

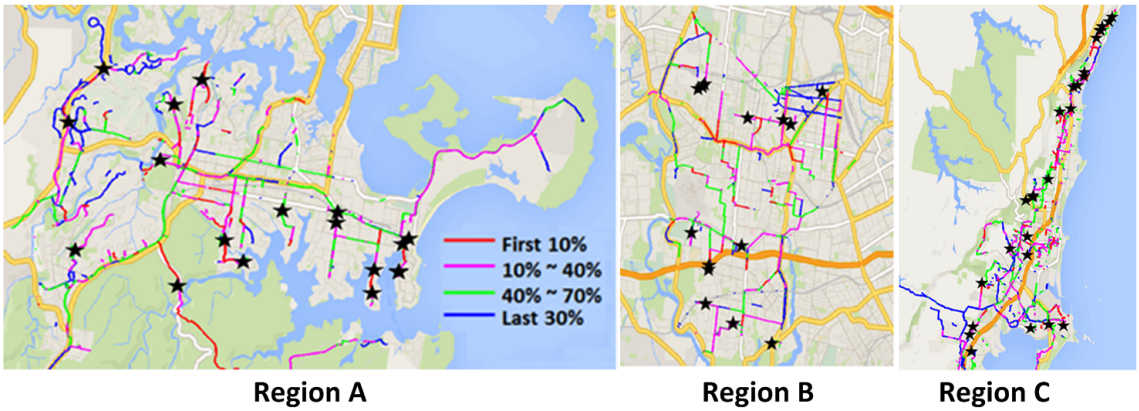

Figure 3.8: Risk maps for the selected three regions

In addition to the comparison studies shown above, we also demonstrate the real-life impact of the proposed method by showing its improvements in its real-world application. Different from the standard performance measurement, domain experts often adopt evaluation criteria that can reflect the constraints encountered in reality. In the context of water pipe failure prediction, as mentioned before, only a small portion of the pipes can be physically inspected each year. Specifically, due to budget constraint, only 1% of the total CWMs can be inspected every year. Therefore, we show the performance curves with 1% of CWMs inspected in Fig. 3.7. AUC and significance test results are also given in Table 3.3 and Table 3.4 for the situation of inspecting 1% of CWMs. As we can see, the proposed approach significantly outperforms the other methods for all the three regions. In region C, the proposed approach nearly doubles the number of detected failures compared with the second best method.

A risk map, as shown in Fig. 3.8, is another widely used method for visualizing real-life impact. As illustrated in the figure, the prioritization of pipes is coded by different colours. For instance, red lines indicate the top 10% high-risk pipes predicted by our method. Black stars in the figure denote the failures which occurred in the testing year. As we can see, many failures could be prevented and significant economic and social savings could be brought to the water utility if the proposed method were applied.

3.6 Conclusions

In this chapter, I present the Dirichlet process mixture of hierarchical beta process model for water pipe failure prediction. The model demonstrates high adaptability to the diversity of failure patterns. Its structure and complexity can grow as the number of data points increases. It tackles the sparse failure data problem by sharing failure data through pipe grouping. An efficient Metropolis-within-Gibbs sampling algorithm is also proposed for handling large-scale datasets. The empirical studies conducted on the real water pipe data verifies the superiority of the proposed approach.

Chapter 4

Bayesian Nonparametric Approach for Event Interaction

The previous chapter aims to solve the sparsity problem in the real-world temporal event modelling. This chapter focuses on the interactions of events. Temporal events are usually not isolated. One event may trigger the occurrence of another. Hence, understanding such interactions can help us better model the generation mechanism of events and thereby providing more accurate prediction. Particularly, in this chapter, a distance dependent prior over branching structure is developed to describe the relationship between events. The proposed model, namely the infinite branching model (IBM), generalizes interaction point processes(IPPs) to model the infinite interaction between events.

4.1 Introduction and Motivation

Most events in the world are generally non-independent, by which it means one event may cause or repel the occurrences of others. Examples can be readily found in various areas. For instance, many biological phenomena compete for local resources, hence demonstrate spatial over-dispersion property. Strong clustering patterns are

often observed by seismologists [11] and epidemiologists [98], as earthquakes and epidemics are well known diffusible events. Buy and sell trades in financial markets also arrive in clusters [99]. Information prorogation in social network shows contagious and clustering trait [100]. 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 [101] provides us a generic yet adaptable tool for modelling series of events occurring at random locations and times. It considers a random collection of points falling in some space. When modelling 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 modelling purposes. In this chapter, we mainly focus on interaction

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

point processes (IPPs) [73, 74] 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 [75, 76], is proposed with the consideration of the aforementioned three tasks.

Many statistical methods exist for modelling events in spatial and temporal space, and most of them make an ex-changeability assumption on a certain component of the overall model [102]. For instance, random walk models [103] and the infinite hidden Markov model [39], which are widely used for discrete time series, assume Markov ex-changeability [104, 105] 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 ex-changeability over increments.

However, in general, the observed events in spatial and temporal space are seldom exchangeable, especially for their dependencies. Distance dependent Chinese Restaurant process (ddCRP) [106] is a simple yet flexible class of distributions over partitions allowing non-exchangeability. It can be used for directly modelling 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).

4.2 Infinite Branching Model

Inspired by the Poisson cluster process, in this work, we propose a Bayesian statistical model, the IBM, that generalises IPPs as a mixture of Poisson processes. The key

component of IBM is a distance dependent prior over branching structure of points. As mentioned in the introduction section, most of the statistical models designed for spatio-temporal events assume ex-changeability, which is unrealistic for modelling point dependencies. Hence, we adapt the ddCRP, a class of non-exchangeable distributions over partitions, as a prior of branching structure.

4.2.1 Prior Belief of Branching Structure

The direct modelling 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:

$$p(c_j = i|\eta, f, D) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \eta & \text{if } i = j, \end{cases} \tag{4.1}$$

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.

Dist-CRP exhibits an interesting 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 2.4.1.2. Thus, the ddCRP can be used as a distribution over branching structures.

4.2.2 Infinite Branching model

With the support of DD-CRP as the prior over branching structure , the Infinite Branching model(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 modelling 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. We use $R(c_{1:N})$ to represent the mapping from point assignment to point cluster assignment, $R^O(c_{1:N})$ to represent the immigrant of the corresponding cluster, and $R^I(c_{1:N})$ to represent the offspring of the corresponding cluster. 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 $\lambda_1^\beta \sim \text{Gamma}(\gamma_1, \gamma_2)$.
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 normalised offspring intensity $\lambda^\beta_{R(c_n)} \sim \text{Gamma}(\gamma_1, \gamma_2)$.

In the above, λ^μ, λ^α, γ_1 and $gamma_2$ are hyper-parameters. $\mathcal{PP}(\cdot)$ indicates a Poisson process. Samples can be drawn from an inhomogeneous Poisson process by utilising a thinning process, a point process variant of rejection sampling. Specifically, the Ogata's modified thinning [107] can be used, as summarised by the Algorithm 7.5.IV in [101].

The graphic model is given by Fig. 4.1

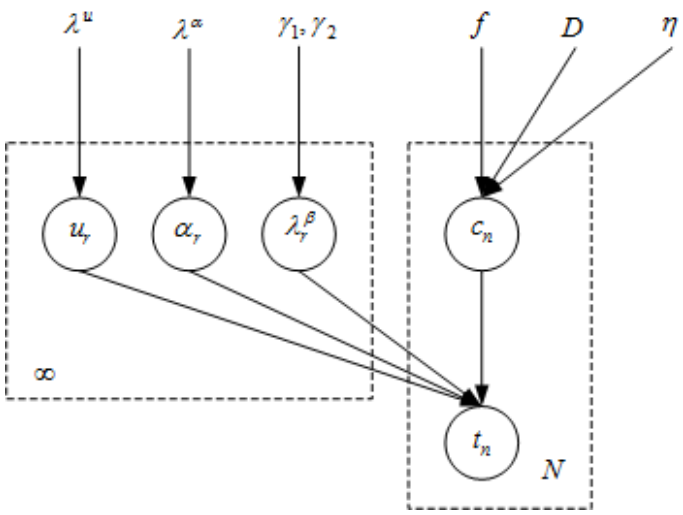

Figure 4.1: Graphic model of infinite branching model

Where, the subscript r denotes the infinite choice of $R(c_{1:N})$.

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 modelling 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.

Such prior fits for some real-world phenomena, *e.g.*, tweets from opinion leaders or celebrities often invoke huge amount of following tweets. The "rich gets richer" behaviour exhibits. In a special case, the branching structure in IBM can be exchangeable. That is when the distance in ddCRP is proportional to the number of descendent, e.g., sequential CRP, the probability of a branching structure depends only on the size of the descendent, but not on which events are children of which. Hence, it becomes exchangeable. However, in general, the branching structure in IBM is not exchangeable. For instance, the influence estimation of social information, influence maximisation, and the probability of ultimate extinction, of the influential idea leaders in LinkedIn, Facebook and Tweeter, anyway it is a prior reflexing your belief of reality, you can design the distance, for instance like the affinity prorogation, both the similarity and the responsibility are considered.

4.2.3 Relation with IRM

The IBM is related to the infinite relational model (IRM) [108]. The IRM aims at inferring meaningful latent structure within observed graph or network. An unbounded number of blocks of nodes with similar behaviour can be automatically revealed with the support of the CRP prior on node partitions. Distinctively, 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.

4.3 Hierarchical Infinite Branching 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 normalised 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 $\mu \sim \text{Exponential}(\lambda^\mu)$.
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 $\lambda_1^\beta \sim \text{Gamma}(\gamma_1, \gamma_2)$.
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 $\alpha_h \sim \text{Exponential}(\lambda^\alpha)$, and sample scale parameter for normalised offspring intensity $\lambda_h^\beta \sim \text{InverseGamma}(\gamma_1, \gamma_2)$.

In the above, λ^{μ}, λ^{α}, γ_1, γ_2 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 [109] 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 hierarchical model can automatically discover the point clusters that share the same triggering scheme even when they are disjoint in spatiotemporal space.

4.4 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 [101], with the support of local Janossy density, the likelihood function for a realization X of a regular point process can be represented as:

$$L = \left(\prod_{i=1}^{N} \lambda(t_i) \right) \exp\left(- \int_0^T \lambda(t) dt \right), \tag{4.2}$$

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 modelling the branching structure grants us the computational simplicity to decompose the likelihood function into independent parts:

$$p(X|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, C) = p(I|\mu, C) \prod_{i=1}^{N} p(O_i|\alpha_{R(c_i)}, \beta_{R(c_i)}, C), \tag{4.3}$$

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:

$$p(I|\cdot) = \frac{\prod_{t_i \in I} \mu(t_i)}{\exp\left(\int_0^T \mu(t)dt\right)}, \tag{4.4}$$

$$p(O_i|\cdot) = \frac{\prod_{t_j \in O_i} \alpha_{R(c_i)}\beta_{R(c_i)}(t_j - t_i)}{\exp\left(\alpha_{R(c_i)} \int_{t_i}^T \beta_{R(c_i)}(t - t_i)dt\right)}. \tag{4.5}$$

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 [34] 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:

$$A_\mu = \frac{p(\hat{\mu})}{p(\mu)} \prod_{t_i \in I} \left(\frac{\hat{\mu}(t_i)}{\mu(t_i)}\right) \exp\left(\int_0^T \mu(t)dt - \int_0^T \hat{\mu}(t)dt\right), \tag{4.6}$$

$$\begin{aligned} A_{\alpha_{R(c_i)}} =& \frac{p(\hat{\alpha}_{R(c_i)})}{p(\alpha_{R(c_i)})} \prod_{t_j \in R'(c_i)} \left(\frac{\hat{\alpha}_{R(c_i)}}{\alpha_{R(c_i)}}\right) \\ & \cdot \exp\left(\sum_{t_j \in R'(c_i)} \left(\alpha_{R(c_i)} - \hat{\alpha}_{R(c_i)}\right) B_{R(c_i)}\right), \end{aligned} \tag{4.7}$$

$$\begin{aligned} A_{\beta_{R(c_i)}} =& \frac{p(\hat{\beta}_{R(c_i)})}{p(\beta_{R(c_i)})} \prod_{t_j \in R'(c_i)} \left(\frac{\hat{\beta}_{R(c_i)}(t_j - t_{c_j})}{\beta_{R(c_i)}(t_j - t_{c_j})}\right) \\ & \cdot \exp\left(\sum_{t_j \in R'(c_i)} \alpha_{R(c_i)}\mathbf{U}\right). \end{aligned} \tag{4.8}$$

For the above Hastings ratios, we assume the prior distributions of parameters are independent. In Eq. 4.7 and Eq. 4.8, 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. 4.1 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:

$$A_{c_i}^{I\to O} = \frac{\alpha_{R(\hat{c}_i)}\beta_{R(\hat{c}_i)}(t_i - t_{\hat{c}_i})}{\mu(t_i)} \cdot \prod_{t_j \in R'(c_i)} \mathbf{V} \cdot \exp\left(\sum_{t_j \in R'(c_i)} \mathbf{W}\right), \tag{4.9}$$

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. 4.9 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:

$$A_{c_i}^{O\to I} = \frac{\mu(t_i)}{\alpha_{R(c_i)}\beta_{R(c_i)}(t_i - t_{c_i})} \cdot \prod_{t_j \in R'(\hat{c}_i)} \mathbf{V} \cdot \exp\left(\sum_{t_j \in R'(\hat{c}_i)} \mathbf{W}\right), \tag{4.10}$$

where $\mathbf{V}$ and $\mathbf{W}$ are as defined before. The second part of Eq. 4.10 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:

$$A_{c_i}^{O\to\hat{O}} = \frac{\alpha_{R(\hat{c}_i)}\beta_{R(\hat{c}_i)}(t_i - t_{\hat{c}_i})}{\alpha_{R(c_i)}\beta_{R(c_i)}(t_i - t_{c_i})} \prod_{t_j \in R'(c_i) \wedge t_j \in R'(\hat{c}_i)} \mathbf{V} \cdot \exp\left(\sum_{t_j \in R'(c_i) \wedge t_j \in R'(\hat{c}_i)} \mathbf{W} \right), \tag{4.11}$$

where $\mathbf{V}$ and $\mathbf{W}$ are as defined before. For the second part of Eq. 4.11, 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.

4.5 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. For the synthetic data, we evaluate and visualize the IBM's performance on offspring intensity and branching structure estimations. For the real-world application, we compared the IBM with several popular IPP based models on water pipeline failure prediction and failure type categorization.

4.5.1 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

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 4.1: Results of Diff and LogLik.

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 [82] 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) [110], model independent stochastic declustering (MISD) [11] and Bayesian inference approach for Hawkes process (BHawk) [12]. The comparison results for Diff and log-likelihood are given in Table 4.1. 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

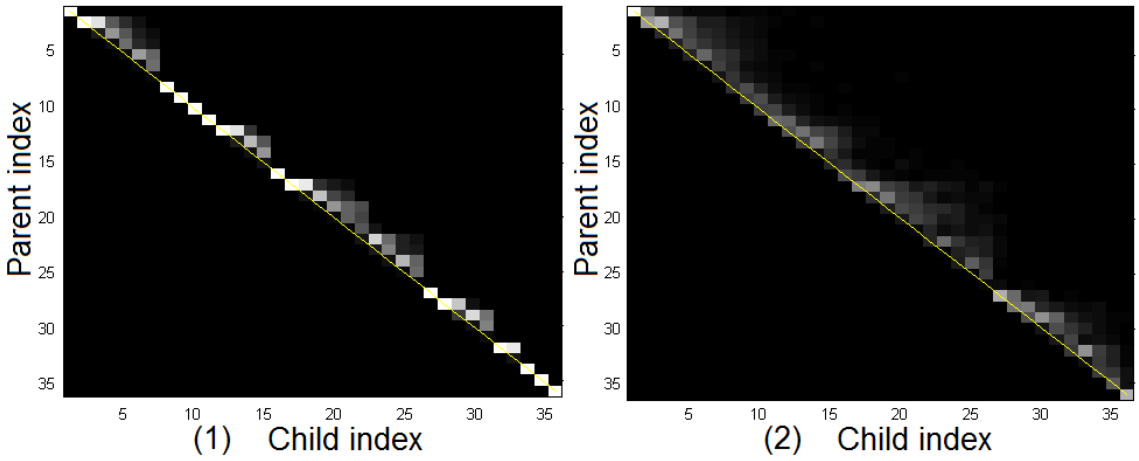

Figure 4.2: Estimated branching structure matrices.

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. 4.2 (1) and Fig. 4.2 (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 colour in the matrices indicates higher probability. As we can see that both matrices show strong clustering behaviours. 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. 4.3 (1) and Fig. 4.3 (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. 4.3 (2), the CRP prior tends to underestimate the number of immigrants and exhibits strong "rich gets richer" behaviour. The ddCRP prior, shown in Fig. 4.3 (1), gives a more accurate estimation for the number of immigrants. Both priors can detect temporal clustering behaviours. The CRP prior tends to find coarser clusters, while the ddCRP prior tends to find finer clusters.

4.5.2 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 behaviours [111, 110, 8, 1, 112]. 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

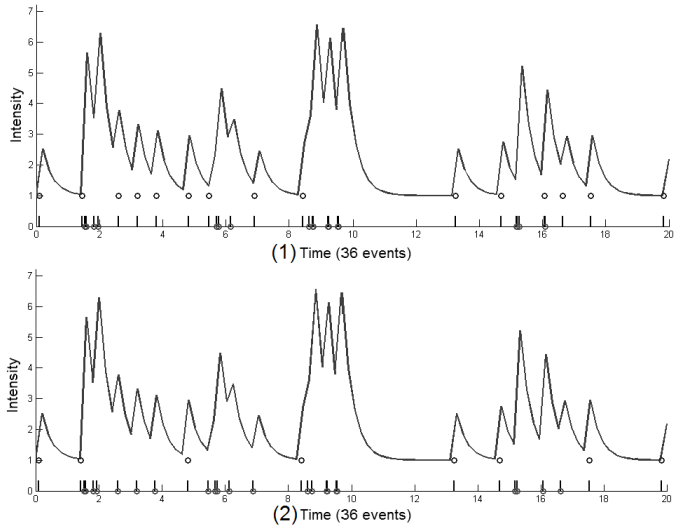

Figure 4.3: Estimated point types.

	# Pipes	Laid years	# Failures	Observation period
District	5121	1930 – 1997	922	2003 – 2010

Table 4.2: Summary of pipe failure data

water utilities to accurately estimate both the type and amount of pipe failures.

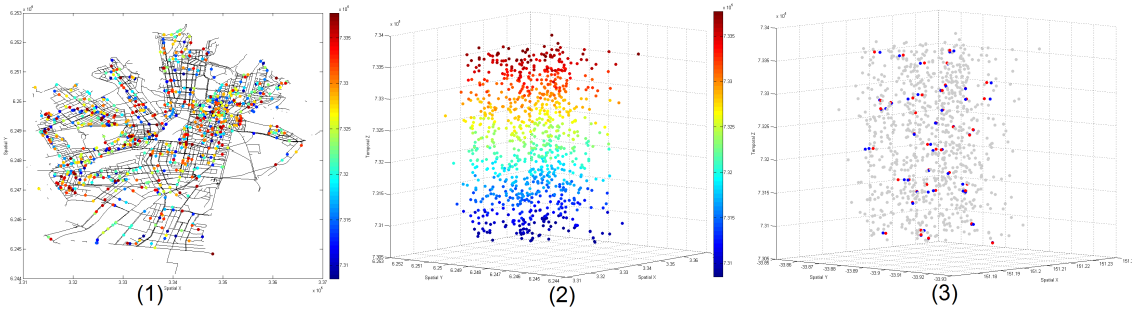

Figure 4.4: Failure points and estimated failure types.

In this experiment, we collected 922 failures from a metropolitan water supply network (The details of the selected dataset are shown in Table 4.2). The failures occurred in a district during 8 years. As shown in Fig.4.4 (1), black lines represent pipelines, warm color dots indicate recent failures and cool color dots indicate old failures. We treat each pipe failure as a point in spatiotemporal space as shown in Fig.4.4 (2). Hence, we can use our model for both failure type estimation and failure

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 4.3: Results of MSE and F1

amount prediction. For the branching structure prior, we adopt a decay function considering both spatial and temporal distances:

$$\begin{aligned} f(d_S, d_T) = I(d_S \leq a_S) \cdot I(d_T \leq a_T) \\ \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{d_S^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{d_T}{\rho}\right), \end{aligned} \tag{4.12}$$

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 modelling 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) [13] and cascades of Poisson process (CPP) [14] 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. The failure type categorization result by the proposed method is also visualized in Fig.4.4 (3). Blue dots indicate the triggered failures and red dots indicate the failures that cause other failures. Additionally, we use silhouette coefficient (SC) [113], 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 4.3. 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.

4.6 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 [114], the connection between branching structure and causality measure of point processes [115] and the extension for multivariate IPPs.

Chapter 5

Bayesian Nonparametric Approach for Event Correlation

In the previous Chapter, we propose a Bayesian statistical model that generalizes and extends interaction point processes to model the interaction of events. In this chapter, we propose a Bayesian nonparametric approach that considers both types of correlations via unifying and generalizing the hidden semi-Markov model and interaction point process model. The proposed approach simultaneously models both the observations and arrival times of temporal events, and automatically determine the number of latent states from data.

5.1 Introduction and Motivation

Temporal events modelling is a classic machine learning problem that has drawn enormous research attentions for decades. It has wide applications in many areas, such as financial modelling, social events analysis, seismological and epidemiological forecasting. An event is often associated with an arrival time and an observation, e.g., a scalar or vector. For example, a trading event in financial market has a trading time and a trading price. A message in social network has a posting time

and a sequence of words. A main task of temporal events modelling is to capture the underlying event correlation and use it to make predictions for future events' observations and/or arrival times.

The correlation between events' observations can be readily found in many real-world cases in which an event's observation is influenced by its predecessors' observations. For examples, the price of a trading event is impacted by former trading prices. The content of a new social message is affected by the contents of the previous messages. State space model (SSM), e.g., the hidden Markov model (HMM) [84], is one of the most prevalent frameworks that consider such correlation. It models the correlation via latent state dependency. Each event in the HMM is associated with a latent state that can emit an observation. A latent state is independent of all but the most recent state, *i.e.*, Markovianity. Hence, a future event observation can be predicted based on the observed events and inferred mechanisms of emission and transition.

Despite its popularity, the HMM lacks the flexibility to model event arrival time. It only allows fixed inter-arrival time. The duration of a type of state follows a geometric distribution with its self-transition probability as the parameter due to the strict Markovian constraint. The hidden semi-Markov model (HSMM) [16, 17] was developed to allow non-geometric state duration. It is an extension of the HMM by allowing the underlying state transition process to be a semi-Markov chain with a variable duration time for each state. In addition to the HMM components, the HSMM models the duration of a state as a random variable and a state can emit a sequence of observations.

The HSMM allows the flexibility of variable inter-arrival times, but it does not consider events' correlation on arrival times. In many real-world applications, one event can trigger the occurrences of others in the near future. For instance, earthquakes and epidemics are diffusible events, *i.e.*, one can cause the occurrences of others. Trading events in financial markets arrive in clusters. Information propa-

gation in social network shows contagious and clustering characteristics. All these events exhibit interaction characteristics in terms of arrival times. The likelihood of an event's arrival time is affected by the previous events' arrival times. Stochastic interaction point process (IPP), *e.g.*, Hawkes process [75], is a widely adopted framework for capturing such arrival time correlation. It models the correlation via a conditional intensity function that depicts the event intensity depending on all the previous events' arrival times. However, unlike the SSMs, it lacks the capability of modelling events' latent states and their interactions.

It is clearly desirable in real-world applications to have both arrival time correlation and observation correlation considered in a unified manner so that we can estimate both when and how events will appear. Inspired by the merits of SSMs and IPPs, we propose a novel Bayesian nonparametric approach that unifies and generalizes SSMs and IPPs via a latent semi-Markov state chain with infinitely countable number of states. The latent states governs both the observation emission and new event triggering mechanism. An efficient sampling method is developed within the framework of particle Markov chain Monte Carlo (PMCMC) [18] for the posterior inference of the proposed model.

5.2 Preliminaries

In this section, we review a technique that is closely related to the proposed method, namely hierarchical Dirichlet process hidden Markov model.

Hierarchical Dirichlet process hidden Markov model (HDP-HMM) [10] is a nice nonparametric extension of HMM. The HDPHMM provides an promising methods to various applications such as natural scene recognition [116] abnormal recognition[117], speaker diarization[118], etc.

In the HDP-HMM model, HDP serves as the prior over the state transitions matrix in order to make the model allow an unbounded set of states. The generative

HDP-HMM model can be summarized as:

$$\begin{aligned} \beta \mid \gamma &\sim \mathrm{GEM}(\gamma), \\ \pi_k \mid \alpha_0, \beta &\sim \mathrm{DP}(\alpha_0, \beta), \\ \theta_k \mid \lambda, H &\sim H(\lambda), \\ s_n \mid s_{n-1}, (\pi_k)_{k=1}^{\infty} &\sim \pi_{s_{n-1}}, \\ y_n \mid s_n, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{s_n}), \end{aligned} \tag{5.1}$$

where, GEM denotes stick-breaking process. The variable sequence π_k indicates the latent state sequence. y_n represents the observation. HDP acts the role of a prior over the infinite transition matrices. Each π_k is a draw from a DP, it depicts the transition distribution from state k. The probability measures from which π_k's are drawn are parameterized by the same discrete base measure β. The emission distribution F is parameterized by θ. Usually H is set to be conjugate of F simplifying inference. γ controls base measure β's degree of concentration. α_0 plays the role of governing the variability of the prior mean measure across the rows of the transition matrix.

As HDP prior does not distinguish self-transitions from transitions to other states thus can not perfectly model the states persistence time, it is vulnerable to unnecessary frequent switching of states and more states. [119] proposed a sticky HDP-HMM to include a self-transition bias parameter into the state transition measure $\pi_k \sim DP(\alpha_0 + \kappa, (\alpha_0\beta + \kappa\delta_k)/(\alpha_0 + \kappa))$, where κ controls the stickiness of the transition matrix. Another remarkable variant to solve this problem is Hierarchical Dirichlet process hidden semi-Markov model (HDP-HSMM) [21] that can learn non-geometric state duration by introducing a super state sequence that governs the states' duration distribution.

5.3 Infinite Hidden Semi-Markov Modulated Interaction Point Process

The HMM has proven to be an excellent general framework modelling sequential data, but it has two significant drawbacks: (1) The durations of events (or the inter-arrival times between events) are fixed to a common value. The state duration distributions are restricted to a geometric form. Such setting lacks the flexibility for real-world applications. (2) The size of the latent state space in the HMM must be set a priori instead of learning from data.

The hidden semi-Markov model (HSMM) [16, 17] is a popular extension to the HMM, which tries to mitigate the first drawback of the HMM. It allows latent states to have variable duration, thereby forming a semi-Markov chain. It reduces to the HMM when duration follow a geometric distribution. Additional to the 4 components of the HMM, HSMM has a state duration probability distribution. As a result, the inference procedure for the HSMM also involves the inference of the duration probability distribution.

The recent development in Bayesian nonparametric theory helps address the second drawback of the HMM. In HDP-HMM model, Because the HMM can be treated as a set of mixture models in a dynamic manner, each of which corresponds to a value of the current state, the HDP becomes a natural choice as the prior over the state transitions [39, 10].

We notice that all the aforementioned SSMs(HMM, HSMM and HDP-HMM) neglect the interaction between events in terms of event arrival time, which is well-depicted in interaction point processes (IPPs). We propose an infinite hidden semi-Markov modulated interaction point process model (iHSMM-IPP), which hybridizes SSMs and IPPs by a natural way and thereby owns the merits of both.

iHSMM-IPP is a Bayesian nonparametric stochastic point process with a latent semi-Markov state chain determining both event emission probabilities and event

triggering kernels. An intuitive illustration is given in Fig. 5.1. Each temporal event in the iHSMM-IPP is represented by a stochastic point and each point is associated with a hidden discrete state $\{s_i\}$ that plays the role of determining event emission and triggering mechanism. As in SSMs and IPPs, the event emission probabilities guide the generation of event observations $\{y_i\}$ and the event triggering kernels influence the occurrence times $\{t_i\}$ of events. The hidden state depends only on the most recent event's state. The size of the latent state space is infinite countable with the HDP prior.

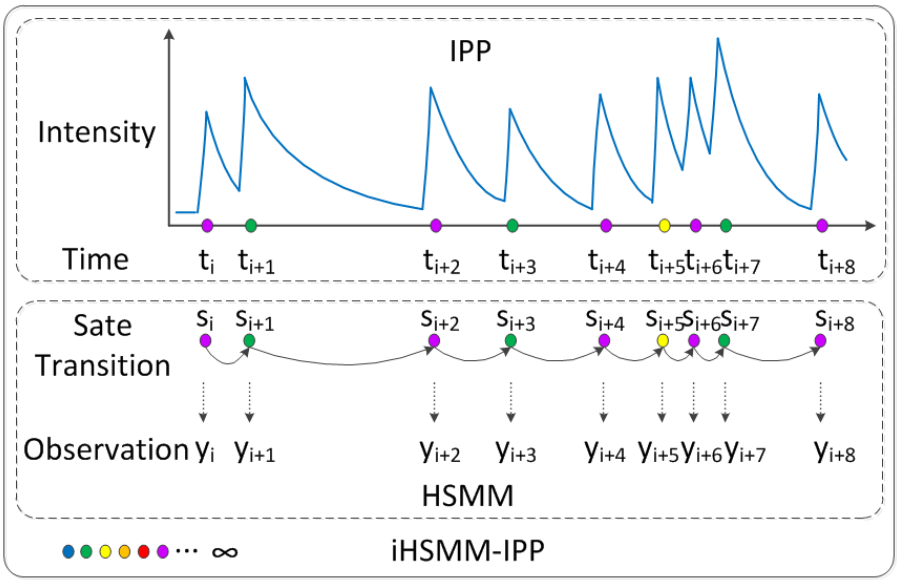

Figure 5.1: An intuitive illustration of the iHSMM-IPP model.

In Fig. 5.1, every event in the iHSMM-IPP model is associated with a latent state s, an arrival time t and an observable value y. The colours of points indicate latent states. Blue curve shows the event intensity. The top part of the figure illustrates the IPP component of the iHSMM-IPP model and the bottom part illustrates the HSMM component. The two components are integrated together via an infinite countable semi-Markov latent state chain.

The model can be formally defined as the following.

$$\begin{aligned} \beta \mid \gamma \backsim \mathrm{GEM}(\gamma), \quad \pi_k \mid \alpha_0, \beta \backsim \mathrm{DP}(\alpha_0, \beta), \quad \theta_k \mid \eta, H \backsim H(\eta), \\ \rho_k \mid \chi, H' \backsim H'(\chi), \quad s_n \mid s_{n-1}, (\pi_k)_{k=1}^{\infty} \backsim \pi_{s_{n-1}}, \\ t_n \mid \cdot \backsim \mathcal{PP}(\mu + \sum_{i=1}^{n-1} \psi_{\rho_{s_i}}(t - t_i)), \quad y_n \mid s_n, (\theta_k)_{k=1}^{\infty} \backsim F(\theta_{s_n}). \end{aligned} \tag{5.2}$$

The corresponding graphical model is given in Fig. 5.2.

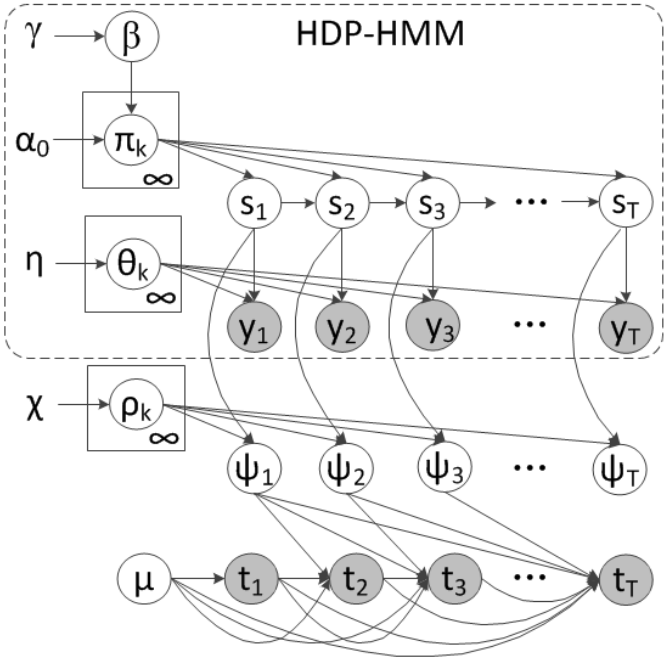

Figure 5.2: Graphical model of the iHSMM-IPP model.

We use $\psi_{\rho_{s_i}}(\cdot)$ to denote the triggering kernel parameterized by ρ_{s_i} which is indexed by latent state s_i. We use $\psi_{s_i}(\cdot)$ instead of $\psi_{\rho_{s_i}}(\cdot)$ for the remaining of the paper for the sake of notation simplicity. The iHSMM-IPP is a generative model that can be used for generating a series of events with arrival times and emitted observations. The arrival time t_n is drawn from a Poisson process. We do not consider edge effect in this work. Therefore, the first event's arrival time, t_1, is drawn from a homogeneous Poisson process parameterized by a hyper-parameter μ.

For $n > 1$, t_n is drawn from an inhomogeneous Poisson process whose conditional intensity function is defined as: $\mu + \sum_{i=1}^{n-1} \psi_{s_i}(t - t_i)$. As defined before, $\psi_{s_i}(\cdot)$ indicates the triggering kernel of a former point i whose latent state is s_i. The state of the point s_n is drawn following the guidance of the HDP prior as in the HDP-HMM. The emitted observation y_n is generation from the emission probability distribution $F(\cdot)$ parameterized by θ_{s_n} which is determined by the state s_n.

5.4 Posterior Inference for iHSMM-IPP

In this section, we describe the inference method for the proposed iHSMM-IPP model. Despite its flexibility, the proposed iHSMM-IPP model faces three challenges for efficient posterior inference: (1) strong correlation nature of its temporal dynamics (2) non-Markovianity introduced by the event triggering mechanism, and (3) infinite dimensional state transition.

The traditional sampling methods for high dimensional probability distributions, *e.g.*, MCMC, sequential Monte Carlo (SMC), are unreliable when highly correlated variables are updated independently, which can be the case for the iHSMM-IPP model. So we develop the inference algorithm within the framework of particle MCMC (PMCMC), a family of inferential methods recently developed in [18]. The key idea of PMCMC is to use SMC to construct a proposal kernel for an MCMC sampler. It not only improves over traditional MCMC methods but also makes Bayesian inference feasible for a large class of statistical models.

For tackling the non-Markovianity, ancestor resampling scheme [19] is incorporated into our inference algorithm. As existing forward-backward sampling methods, ancestor resampling uses backward sampling to improve the mixing of PMCMC. However, it achieves the same effect in a single forward sweep instead of using separate forward and backward sweeps. More importantly, it provides an effective way of sampling for non-Markovian SSMs.

5.4.1 Particle Gibbs Sampling for iHSMM-IPP

Given a sequence of N events, $\{y_n, t_n\}_{n=1}^N$, the inference algorithm needs to sample the hidden state sequence, $\{s_n\}_{n=1}^N$, emission distribution parameters $\theta_{1:K}$, background event intensity μ, triggering kernel parameters, $\psi_{1:K}$ (we omit ρ and use $\psi_{1:K}$ instead of $\psi_{\rho_{1:K}}$ for notation simplicity as before), transition matrix, $\boldsymbol{\pi}_{1:K}$, and the HDP parameters (α_0, γ, κ, β). We use K to represent the number of active states and Ω to indicate the set of variables excluding the latent state sequence, *i.e.*, $\Omega = \{\alpha_0, \beta, \gamma, \kappa, \mu, \theta_{1:K}, \psi_{1:K}, \pi_{1:K}\}$. Only major variables are listed, and Ω may also include other variables, such as the probability of initial latent state. At a high level, all the variables are updated iteratively using a particle Gibbs (PG) sampler. A conditional SMC is performed as a proposal kernel for updating latent state sequence in each PG iteration. An ancestor resampling scheme is adopted in the conditional SMC for handling the non-Markovianity caused by the triggering mechanism. Metropolis sampling is used in each PG iteration to update background event intensity μ and triggering kernel parameters $\psi_{1:K}$. The remaining variables in Ω can be sampled by following the scheme in [119, 10] readily. The proposal distribution $q_\Omega(\cdot)$ in the conditional SMC can be set by following [20]. The PG sampler is given in the following:

Step 1: Initialization, $i = 0$, set $\Omega(0)$, $s_{1:N}(0)$, $B_{1:N}(0)$.

Step 2: For iteration $i \geqslant 1$

- **(a)** Sample $\Omega(i) \sim p\{\cdot|y_{1:N}, t_{1:N}, s_{1:N}(i-1)\}$.
- **(b)** Run a conditional SMC algorithm targeting $p_{\Omega(i)}(s_{1:N}|y_{1:N}, t_{1:N})$ conditional on $s_{1:N}(i-1)$ and $B_{1:N}(i-1)$.
- **(c)** Sample $s_{1:N}(i) \sim \hat{p}_{\Omega(i)}(\cdot|y_{1:N}, t_{1:N})$.

We use $B_{1:N}$ to represent the ancestral lineage of the prespecified state path $s_{1:N}$ and $\hat{p}_{\Omega(i)}(\cdot|y_{1:N})$ to represent the particle approximation of $p_{\Omega(i)}(\cdot|y_{1:N})$. The details

of the conditional SMC algorithm are given in the following. It is worth noting that the conditioned latent state path is only updated via the ancestor resampling.

Step 1: Let $s_{1:N} = \{s_1^{B_1}, s_2^{B_2}, \cdots, s_N^{B_N}\}$ denote the path that is associated with the ancestral lineage $B_{1:N}$

Step 2: For $n = 1$,

(a) For $j \neq B_1$, sample $s_1^j \sim q_\Omega(\cdot|y_1)$, $j \in [1, \cdots, J]$. (J denotes the number of particles.)

(b) Compute weights $w_1(s_1^j) = p(s_1^j)F(y_1|s_1^j)/q_\Omega(s_1^j|y_1)$ and normalize the weights $W_1^j = w_1(s_1^j)/\sum_{m=1}^J w_1(s_1^m)$. (We use $p(s_1^j)$ to represent the probability of the initial latent state and $q_\Omega(s_1^j|y_1)$ to represent the proposal distribution conditional on the variable set Ω.)

Step 3: For $n = 2, \cdots, N$:

(a) For $j \neq B_n$, sample ancestor index of particle j: $a_{n-1}^j \sim \mathcal{C}at(\cdot|W_{n-1}^{1:J})$.

(b) For $j \neq B_n$, sample $s_n^j \sim q_\Omega(\cdot|y_n, s_{n-1}^{a_{n-1}^j})$. If $s_n^j = K+1$ then create a new state using the stick-breaking construction for the HDP:

(i) Sample a new transition probability $\pi_{K+1} \sim \mathcal{D}ir(\alpha_0\beta)$.

(ii) Use stick-breaking construction to expand $\beta \leftarrow [\beta, \beta_{K+1}]$:

$$\beta'_{K+1} \sim \text{Beta}(1, \gamma), \qquad \beta_{K+1} = \beta'_{K+1}\prod_{l=1}^{K}(1-\beta'_l).$$

(iii) Expand transition probability vectors π_k to include transitions to state $K+1$ via the HDP stick-breaking construction:

$$\pi_k \leftarrow [\pi_{k,1}, \cdots, \pi_{k,K+1}], \quad \forall k \in [1, K+1], \text{where}$$

$$\pi'_{k,K+1} \sim \text{Beta}(\alpha_0\beta_{K+1}, \alpha_0(1-\sum_{l=1}^{K+1}\beta_l)), \quad \pi_{k,K+1} = \pi'_{k,K+1}\prod_{l=1}^{K}(1-\pi'_{k,l}).$$

(iv) Sample parameters for a new emission probability and triggering kernel $\theta_{K+1} \sim H$ and $\psi_{1:K} \sim H'$.

(d) Perform ancestor resampling for the conditioned state path. Compute the ancestor weights $\tilde{w}^{p,j}_{n-1|N}$ via Eq. 5.5 and Eq. 5.6 and resample $a_n^{B_n}$ as $p(a_n^{B_n} = j) \propto \tilde{w}^{p,j}_{n-1|N}$.

(e) Compute and normalize particle weights:

$$w_n(s_n^j) = \pi(s_n^j | s_{n-1}^{a_{n-1}^j}) F(y_n|s_n^j)/q_\Omega(s_n^j|s_{n-1}^{a_{n-1}^j}, y_n),\ W_n(s_n^j) = w_n(s_n^j)/(\sum_{j=1}^{J} w_n(s_n^j)).$$

5.4.2 Metropolis Sampling for Background Intensity and Triggering Kernel

For the inference of the background intensity μ and the parameters of triggering kernels ψ_k in the step 2 (a) of the PG sampler, Metropolis sampling is used. As described in [72], the conditional likelihood of the occurrences of a sequence of events in IPP can be expressed as:

$$\mathcal{L} \triangleq p(t_{1:N}|\mu, \psi_{1:K}) = \left(\prod_{n=1}^{N} \lambda(t_n)\right) \exp\left(-\int_0^T \lambda(t)dt\right). \tag{5.3}$$

We describe the Metropolis update for ψ_k, and similar update can be derived for μ. The normal distribution with the current value of ψ_k as mean is used as the proposal distribution. The proposed candidate ψ_k^* will be accepted with the probability: $A(\psi_k^*, \psi_k) = \min\left(1, \frac{\hat{p}(\psi_k^*)}{\hat{p}(\psi_k)}\right)$. The ratio can be computed as:

$$\begin{aligned}\frac{\hat{p}(\psi_k^*)}{\hat{p}(\psi_k)} &= \frac{p(\psi_k^*)}{p(\psi_k)} \cdot \frac{p(t_{1:N}|\psi_k^*, \text{rest})}{p(t_{1:N}|\psi_k, \text{rest})} = \frac{p(\psi_k^*)}{p(\psi_k)} \cdot \left(\prod_{n=1}^{N} \frac{\mu(t_n) + \sum_{u<n} \psi_{s_u}^*(t_n - t_u)}{\mu(t_n) + \sum_{u<n} \psi_{s_u}(t_n - t_u)}\right) \\ &\cdot \exp\left(\sum_{u\in[1,N]} \left(\Psi_{s_u}(T - t_u) - \Psi_{s_u}^*(T - t_u)\right)\right).\end{aligned} \tag{5.4}$$

We use $\Psi(\cdot)$ to represent the cumulative distribution function of the kernel function $\psi(\cdot)$. We use $\psi_{s_u}^*(\cdot)$ to represent the u-th event's triggering kernel candidate if

$s_u = k$. It remains the current triggering kernel otherwise. $[0, T]$ indicates the time period of the N events.

5.4.3 Truncated Ancestor Resampling for Non-Markovianity

Truncated ancestor resampling [19] is used for tackling the non-Markovianity caused by the triggering mechanism of the proposed model. The ancestor weight can be computed as:

$$\tilde{w}^{p,j}_{n|N} = w^j_n \frac{\gamma_{n+p}(\{s^j_{1:n}, s'_{n+1:n+p}\})}{\gamma_n(s^j_{1:n})} \tag{5.5}$$

$$\frac{\gamma_{n+p}(\{s^j_{1:n}, s'_{n+1:n+p}\})}{\gamma_t(s^j_{1:n})} = \frac{p(s_{1:p}, y_{1:p}, t_{1:p})}{p(s_{1:n}, y_{1:n}, t_{1:n})} = \frac{\mathcal{L}(t_{1:p})}{\mathcal{L}(t_{1:n})} \cdot \prod_{j=n+1}^{p} F(y_j|s_j)\pi(s_j|s_{j-1}) \tag{5.6}$$

For notation simplicity, we use w^j_n to represent $w_n(s^j_n)$. In general, $n + p$ needs to reach the last event in the sequence. However, due the computational cost and the influence decay of the past events in the proposed iHSMM-IPP, it is practical and feasible to use only a small number of events as an approximation instead of using all the remaining events in Eq. 5.6.

5.5 Empirical Study

In the following experiments, we demonstrate the performance of the proposed inference algorithm based on synthetic data and show the applications of the proposed iHSMM-IPP model in real-world settings.

5.5.1 Synthetic Data

As in [120, 119, 20], we generate the synthetic data of 1000 events via a 4-state Gaussian emission HMM with self-transition probability of 0.75 and the remaining probability mass uniformly distributed over the other 3 states. The means of emission are set to -2.0 -0.5 1.0 4.0 with the deviation of 0.5. The occurrence times of

	Synthetic
Method	Diff
iHSMM-IPP	0.36
M-MHawkes	0.55
VI-MHawkes	0.62

Table 5.1: Results on Synthetic data

events are generated via the Hawkes process with 4 different triggering kernels, each of which corresponds to a HMM state. The background intensity is set to 0.6 and the triggering kernels take the exponential form: $\lambda(t) = 0.6 + \sum_{t_n<t} \alpha' \cdot \exp(-\beta'(t-t_n))$ with $\{0.1, 0.9\}, \{0.5, 0.9\}, \{0.1, 0.6\}, \{0.5, 0.6\}$ as the $\{\alpha', \beta'\}$ parameter pairs of the kernels. A thinning process [107] (a point process variant of rejection sampling) is used to generate event times of Hawkes process.

We compared 4 related methods to demonstrate the performance of the proposed iHSMM-IPP model and inference algorithm: particle Gibbs sampler for sticky HDP-HMM [20], weak-limit sampler for HDP-HSMM [21], Metropolis-within-Gibbs sampler for marked Hawkes process [12] and variational inference for marked Hawkes process [121]. The normalized Hamming distance error is used to measure the performance of the estimated state sequences. The Diff distance used in [82] (i.e., $\frac{\int(\tilde{\psi}(t)-\psi(t))^2 dt}{\int(\psi(t))^2 dt}$, $\psi(t)$ and $\tilde{\psi}(t)$ represent the true and estimated kernels respectively) is adopted for measuring the performance of the estimated triggering kernels. The estimated ones are greedily matched to minimize their distances from the ground truth.

The average results of the normalized Hamming distance errors are shown in Fig. 5.3 and the Diff distance errors are shown in the second column of Table 5.1. The results show that the proposed inference method can not only quickly converge to an accurate estimation of the latent state sequence but also well recover the underlying triggering kernels. Its clear advantage over the compared SSMs and marked Hawkes processes is due to its considerations of both occurrence times and emitted observations for the inference.

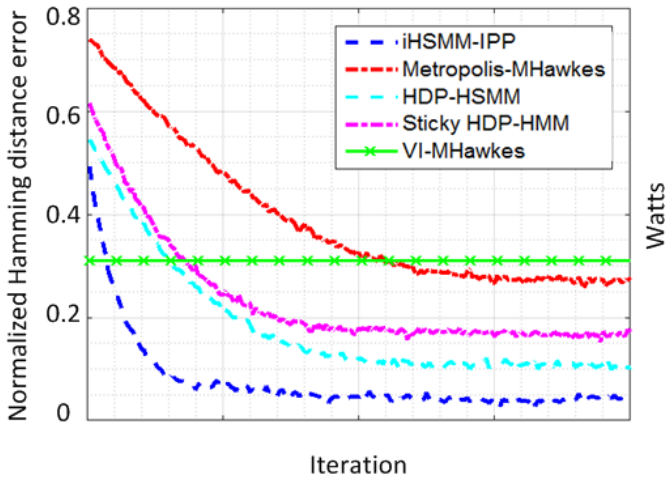

Figure 5.3: Normalized Hamming distance errors for synthetic data.

5.5.2 Understanding Energy Consumption Behaviours of Households

In this section, we use energy consumption data from the Reference Energy Disaggregation Dataset (REDD) [122] to demonstrate the application of the proposed model. The dataset was collected via smart meters recording detailed appliance-level electricity consumption information from approximately 40 homes in the Boston and San Francisco metropolitan areas. All of the data were collected for 48 different circuit breakers, with the collection period for each home typically ranging from 2 to 4 weeks. The dataset was collected with the intension to understand household energy usage patterns and make recommendations for efficient consumption. The 1 Hz low frequency REDD data is used and down sampled to 1 reading per minute covering 1 day energy consumption. Very low and high consumption readings are removed from the reading sequence. Fig. 5.4 $(Left)$ shows the cleaned reading sequence. Readings are in Watts and colours indicate appliance types: lighting, refrigerator, disposal, dishwasher, washer dryer, kitchen outlets, microwave, stove.

The appliance types are modelled as latent states in the proposed iHSMM-IPP model. The readings are the emitted observations of states governed by Gaussian

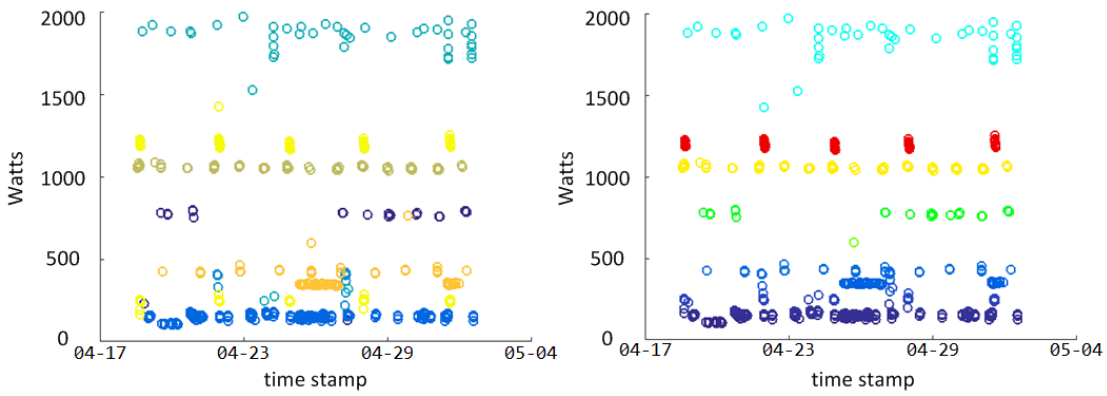

Figure 5.4: Left: Cleaned energy consumption readings of the REDD dataset. Right: Estimated states by the proposed iHSMM-IPP model.

distributions. The relationship between the usages of different appliances is modelled via the state transition matrix. The triggering kernels of states in the model depict the influences of appliances on triggering the following energy consumption, e.g., the usage of washing machine triggers the following energy usage of dryer. As in the first experiment, exponential form of trigger is adopted and independent exponential priors with hyper-parameter 0.01 are used for kernel parameters (α', β'). Fig. 5.4 (*Right*) shows the estimated states by the proposed iHSMM-IPP model.

The 4 methods used in the first experiment are compared with the proposed model. The average results of the normalized Hamming distance errors and the log likelihoods are shown in the third and fourth columns of Table 5.2. The proposed model outperforms the other methods due to the fact that it has the flexibility to capture the interaction between the usages of different appliances. Other models mainly rely on the emitted observations, *i.e.*, readings for inferring the types of appliances.

	REDD	
Method	Hamming	LogLik
iHSMM-IPP	0.30	−120.11
M-MHawkes	0.63	−173.36
VI-MHawkes	0.76	−193.62
HDP-HSMM	0.42	−147.52
S-HDP-HMM	0.55	−163.28

Table 5.2: Results on REDD sets.

5.5.3 Understanding Infrastructure Failure Behaviours and Impacts

Drinking water pipe networks are valuable infrastructure assets. Their failures (e.g., pipe bursts and leaks) can cause tremendous social and economic costs. Hence, it is of significant importance to understand the behaviours of pipe failures (i.e., occurrence time, failure type, labour hours for repair). In particular, the relationship between the types of two consecutive failures, the triggering effect of a failure on the intensity of future failures and the labour hours taken for a certain type of failure can help provide not only insights but also guidance to make informed maintenance strategies.

In this experiment, a sequence of 1600 failures occurred in the same zone within 15 years with 10 different failure types [2] are used for testing the performance of the proposed iHSMM-IPP model (The details of the selected dataset are shown in Table 5.3). Failure types are modelled as latent states. Labour hours for repair are emissions of states, which are modelled by Gaussian distributions. It is well observed in industry that pipe failures occur in clusters, i.e., certain types of failures can cause high failure risk in near future. Such behaviours are modelled via the triggering kernels of states.

	# Pipes	Laid years	# Failures	Observation period
District	12461	1930 − 1997	1600	1998 − 2012

Table 5.3: Summary of pipe failure data

	Pipe			
Method	Hamming	LogLik	MSE Failures	MSE Hours
iHSMM-IPP	0.39	−677	82.8	28.6
M-MHawkes	0.64	−1035	142.2	80.2
VI-MHawkes	0.78	−1200	166.7	93.7
HDP-HSMM	0.52	−850	103.8	42.3
S-HDP-HMM	0.59	−993	128.5	55.9

Table 5.4: Results of the water pipe dataset.

As in the first experiment, we compare the proposed iHSMM-IPP model with 4 related methods. The sequence is divided into two parts 90% and 10%. The first part of the sequence is used for training models. The normalized Hamming distance errors and log likelihoods are used for measuring the performances on the first part. Then the models are used for predicting the remaining 10% of the sequence. The predicted total number of failures and total labour hours for each failure type are compared with ground truth by using mean square error. The results are shown in the last four columns of Table 5.4. It can be seen that the proposed iHSMM-IPP achieves the best performance for both the estimation on the first part of the sequence and the prediction on the second part of the sequence. Its superiority comes from the fact that it well utilizes both the observed labour hours and occurrence times while others only consider part of the observed information or have limitations on model flexibility.

5.6 Conclusion

In this work, we proposed a new Bayesian nonparametric stochastic point process model, namely the infinite hidden semi-Markov modulated interaction point process model. It captures both emitted observations and arrival times of temporal events for capturing the underlying event correlation. A Metropolis-within-particle-Gibbs sampler with truncated ancestor resampling is developed for the posterior inference

of the proposed model. The effectiveness of the sampler is shown on a synthetic dataset with the comparison of 4 related state-of-the-art methods. The superiority of the proposed model over the compared methods is also demonstrated in two real-world applications, *i.e.*, household energy consumption modelling and infrastructure failure modelling.

Chapter 6

Conclusion and Future Work

In the previous chapters, a set of novel BNP models are developed to learn the underlying relations of stochastic events series. The proposed BNP models allow us to extract the underlying complex patterns of observed data with fewer assumptions about the model form. In this chapter, I summarize the main contributions and discuss potential future directions.

6.1 Works Summary

In general, this thesis successfully demonstrates that BNP theory serves as a promising pathway to introduce flexibility and extendibility in modelling the complicated relationships and structures for stochastic temporal events. Specifically, I have made several advances as listed in the following.

Tackling Data Sparsity: Data sparsity problem has been one of the most common challenges in machine learning applications. The true underlying structure of patterns and relations are difficult to capture as there is not enough data to train the model. In Chapter 3, I demonstrated that flexibly clustering capability granted by Dirichlet process helps alleviate the data sparsity problem in water pipes failure prediction application. The proposed model, namely Dirichlet process mixture of Hi-

erarchical beta process (DPMHBP), makes it possible to share failure data among pipes that have similar behaviours to enhance failure prediction. The clustering process is fully data-driven and does not require predefining the number of clusters. The application based on the metropolitan water supply network data has shown the advantage of the proposed model in comparison with other failure prediction models.

Capturing Infinite Interaction: Interaction relation in social and natural events series usually exhibits a cluster property via a specific branching structure. To explore such interaction, in Chapter 4, I established an unexplored theoretical bridge between distance dependent CRP and interaction point processes (IPPs). An integrated model, infinite branching model (IBM), is constructed to estimate point event intensity, interaction mechanism and branching structure simultaneously. The proposed model showed that stochastic events series can be readily represented as infinite branches of points. The prior belief over events' interactions is depicted as the points connections determined by distance, thereby enable the interaction point process to incorporate the observation information in feature space. The empirical study on synthetic data regression and water pipe failure understanding has shown an inspired superiority comparing with other traditional interactive point process models.

Modelling Event Correlation: The events in stochastic event series are correlated both between arrival times and between observations. Separately, interaction point processes (IPPs) concentrate on modelling the former and Hidden Markov model (HMM) focus on capturing the latter. In Chapter 5, I unified and generalized HMM and IPPs via a novel Bayesian nonparametric point process model, which allows a stochastic point process to capture both emitted observations and arrival times of events. The proposed model exploits two types of underlying correlation in a well-integrated way rather than individually. It not only provides us with better prediction accuracy but also allows us to obtain a deeper insight into the interaction mechanism and correlation between events. Several challenges in inference

procedure are also tackled. Experiments based on real world data demonstrated its high potential for event prediction in various domains.

6.2 Suggestion for Future Direction

The result in this thesis has enriched previous insights on stochastic temporal series of natural and social events. Moreover, it has led to several future directions, which can be summarized in 3 aspects as follows.

6.2.1 Computationally Efficient Inference

The inference methods used in this thesis are all based on MCMC framework and have been elaborately optimized. However, due to the potential high geometry of models, for large dataset, the mixing rate can be slow and the convergence diagnosis can be difficult, and computation can be one of the major bottlenecks for system application. This means the proposed models' productivity would be confined for large scale real-world applications. It is desirable to pay enough attention to develop less computationally consuming inference methods for proposed model.

Variational inference[123] technique supplies a sound alternative to MCMC in the context of large-scale problem. It formulates the conditional probability inference as an optimization problem and seeks an exact analytical solution for approximation of the probability. Usually, the optimization goal is set to minimize the Kullbeck-Leibler (KL) distance between the variational distribution and the original distribution. Because the solution of variational inference is deterministic rather than stochastic, it delivers faster solution and has a remarkable advantages on computation resource consumption compared with MCMC.

Considering MCMC's good property of asymptotically exactness, developing a hybrid inference method that makes the best of both will be desirable. Some pre-

vious works such as variational inference for DPMM [124], collapsed variational inference for HDP [125] and mean-field approximation for Hawkes process [126] have supply valuable basis for solving models in current thesis. However, a sophisticated variational approximation for proposed models can still introduce new theoretical challenges due to the inhomogeneity and complicacy of models.

6.2.2 Latent Feature Relation

The thesis dedicates to using Dirichlet process and its extensions to model the cluster relation in stochastic events. In this kind of relation, each event is associated with one single latent cluster. However, in stochastic events series, there is another type of relation in which each observation is associated with several latent features and can be generated based on a distribution parameterized by these features. Thus, how to capture these underlying features in a series of stochastic events is worthwhile to study.

As surveyed in 2.3.3, beta process (BP) can be the model of choice for this task. BP has been successfully applied in infinite latent feature models [127], where a binary matrix with infinite columns generated by BP controls which features are possessed by observations. Similarly, BP can be used to model the allocation of features that each event is associated with. A promising first step is to extend the previous works [128, 129] that tried to using BP to capture the underlying latent dynamical behaviours. The next step that further explore how to model the interaction between events which share the underlying features will be challenging but interesting.

There are also many others relation structures such as nesting structure, blocks structure and graphical structure are worthy further exploring.

6.2.3 Applications

I have proposed innovative Bayesian nonparametric approaches for modelling stochastic temporal events. It is expected that the results of this thesis could have wider spectrum of applications in infrastructure, financial and Internet domains.

A potential application is to use Bayesian nonparametric models to learn patterns on financial transaction such as buy and sell intensities of stocks. Due to the large number and diversity of economic indicators, the estimation of invisible states that control the behaviour of market transaction is usually complicated. However, It has been known that market transaction are generally clustered over time and strong characteristics of interaction between transactions are exhibited. It can be imagined that using proposed model in chapter 5 to investigate the transaction events will be predictably fruitful.

Another profitable application is discovery of social clusters. For example, the Twitter and Facebook are a rich source of high quality data as users register with personal information. Each social cluster of users is essentially composed of infinite connections between users. The users' posts have observed a marked self-exciting property and infinite branching structure. Not only the user profile information, but also the temporal information of posts provides important clues to the clustering. The model proposed in chapter 4 supplies an extensible framework to explore the social communities behind the posts. It will be a powerful tool to predict the posts cascade after the occurrence of public accident news.

Bibliography

[1] Peng Lin, Bang Zhang, Yi Wang, Zhidong Li, Bin Li, Yang Wang, and Fang Chen, "Data Driven Water Pipe Failure Prediction: A Bayesian Nonparametric Approach," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, New York, NY, USA, 2015, CIKM '15, pp. 193–202, ACM.

[2] Peng Lin, Bang Zhang, Ting Guo, Yang Wang, and Fang Chen, "Interaction Point Processes via Infinite Branching Model," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, 2016, AAAI'16, pp. 1853–1859, AAAI Press.

[3] Peng Lin, Bang Zhang, Ting Guo, Yang Wang, and Fang Chen, "Infinite Hidden Semi-Markov Modulated Interaction Point Process," in *Advances in Neural Information Processing Systems 29*, pp. 3900–3908. 2016.

[4] Thomas S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, Mar. 1973.

[5] CE Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.," *The Annals of Statistics*, vol. 2, pp. 1152–1174, Nov. 1974.

[6] Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha, *Bayesian Survival Analysis*, Wiley Online Library.

[7] DR COX, "Regression models and life tables," *J Roy Statist Soc B*, vol. 34, pp. 187–220, 1972.

[8] Zhidong Li, Bang Zhang, Yang Wang, Fang Chen, Ronnie Taib, Vicky Whiffin, and Yi Wang, "Water Pipe Condition Assessment: A Hierarchical Beta Process Approach for Sparse Incident Data," *Mach. Learn.*, vol. 95, no. 1, pp. 11–26, Apr. 2014.

[9] David M. Blei and Peter Frazier, "Distance dependent Chinese restaurant processes," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2010*, 2010, pp. 87–94.

[10] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[11] David Marsan and Olivier Lengline, "Extending earthquakes' reach through cascading," *Science*, vol. 319, no. 5866, pp. 1076–1079, 2008.

[12] Jakob Gulddahl Rasmussen, "Bayesian inference for Hawkes processes," *Methodology and Computing in Applied Probability*, vol. 15, no. 3, pp. 623–642, 2013.

[13] Ryan Prescott Adams, Iain Murray, and David J. C. MacKay, "Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 9–16, ACM.

[14] Aleksandr Simma and Michael I Jordan, "Modeling events with cascades of Poisson processes," *arXiv preprint arXiv:1203.3516*, 2012.

[15] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Func-

tions of Markov Chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, Feb. 1970.

[16] Kevin P Murphy, "Hidden semi-markov models (hsmms)," 2002.

[17] Shun-Zheng Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, pp. 215–243, 2010.

[18] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein, "Particle markov chain monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

[19] Fredrik Lindsten, Michael I Jordan, and Thomas B Schn, "Particle gibbs with ancestor sampling.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2145–2184, 2014.

[20] Nilesh Tripuraneni, Shixiang Gu, Hong Ge, and Zoubin Ghahramani, "Particle gibbs for infinite hidden markov models," in *Advances in Neural Information Processing Systems*, 2015, pp. 2395–2403.

[21] Matthew Johnson and Alan S. Willsky, "The Hierarchical Dirichlet Process Hidden Semi-Markov Model," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, 2010, pp. 252–259.

[22] Wasserman Larry, *All of Nonparametric Statistics — Larry Wasserman — Springer*, 2006.

[23] Michael I. Jordan, "Machine Learning from a Nonparametric Bayesian Point of View," 2008.

[24] Nils Lid Hjort, "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data," *The Annals of Statistics*, vol. 18, no. 3, pp. 1259–1294, Sept. 1990.

[25] Oksana Yakhnenko and Vasant Honavar, "Annotating images and image objects using a hierarchical dirichlet process model," in *Proceedings of the 9th International Workshop on Multimedia Data Mining: Held in Conjunction with the ACM SIGKDD 2008*. 2008, pp. 1–7, ACM.

[26] Antonio Torralba, Alan S. Willsky, Erik B. Sudderth, and William T. Freeman, "Describing Visual Scenes using Transformed Dirichlet Processes," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schlkopf, and J. C. Platt, Eds. 2006, pp. 1297–1304, MIT Press.

[27] David M Blei, Perry R Cook, and Matthew Hoffman, "Bayesian nonparametric matrix factorization for recorded music," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 439–446.

[28] Lu Ren, David Dunson, Scott Lindroth, and Lawrence Carin, "Dynamic nonparametric Bayesian models for analysis of music," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 458–472, 2010.

[29] Zhao Xu, Volker Tresp, Achim Rettinger, and Kristian Kersting, "Social network mining with nonparametric relational models," in *Advances in Social Network Mining and Analysis*, pp. 77–96. Springer, 2010.

[30] Percy Liang, Slav Petrov, Michael I Jordan, and Dan Klein, "The Infinite PCFG Using Hierarchical Dirichlet Processes.," in *EMNLP-CoNLL*, 2007, pp. 688–697.

[31] Philip J Cowans, "Information retrieval using hierarchical Dirichlet processes," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2004, pp. 564–565, ACM.

[32] Jayaram Sethuraman, "A constructive definition of Dirichlet priors," *Statistica sinica*, pp. 639–650, 1994.

[33] Yee Whye Teh, "Dirichlet Process," Tech. Rep., Springer US, 2011.

[34] Radford M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*, vol. 9, no. 2, pp. 249–265, 2000.

[35] Percy Liang, Michael I. Jordan, and Ben Taskar, "A permutation-augmented sampler for DP mixture models," in *Proceedings of the 24th International Conference on Machine Learning (ICML-07)*, 2007, pp. 545–552.

[36] Scott Niekum and Andrew G. Barto, "Clustering via Dirichlet Process Mixture Models for Portable Skill Discovery," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., pp. 1818–1826. Curran Associates, Inc., 2011.

[37] Ke Jiang, Brian Kulis, and Michael I. Jordan, "Small-Variance Asymptotics for Exponential Family Dirichlet Process Mixture Models," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 3158–3166. Curran Associates, Inc., 2012.

[38] Konstantina Palla, Zoubin Ghahramani, and David A. Knowles, "A non-parametric variable clustering model," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 2987–2995. Curran Associates, Inc., 2012.

[39] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen, "The infinite hidden Markov model," in *Advances in Neural Information Processing Systems*, 2001, pp. 577–584.

[40] Jun Zhu, Ning Chen, and Eric P. Xing, "Infinite SVM: A Dirichlet Process Mixture of Large-margin Kernel Machines," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 2011, pp. 617–624.

[41] Brian Kulis and Michael I. Jordan, "Revisiting k-means: New Algorithms via Bayesian Nonparametrics," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[42] Lauren Hannah, David M. Blei, and Warren B. Powell, "Dirichlet Process Mixtures of Generalized Linear Models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, 2010, pp. 313–320.

[43] Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei, "Hierarchical Topic Models and the Nested Chinese Restaurant Process," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schlkopf, Eds. 2004, pp. 17–24, MIT Press.

[44] Dae Il Kim, Michael C. Hughes, and Erik B. Sudderth, "The Nonparametric Metadata Dependent Relational Model," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[45] Dongwoo Kim and Alice H. Oh, "Hierarchical Dirichlet Scaling Process," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 973–981.

[46] Drausin Wulsin, Shane Jensen, and Brian Litt, "A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[47] Il M Park, Evan W Archer, Kenneth Latimer, and Jonathan W Pillow, "Universal models for binary spike patterns using centered Dirichlet processes," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges,

L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 2463–2471. Curran Associates, Inc., 2013.

[48] Abel Rodrguez, David B Dunson, and Alan E Gelfand, "The Nested Dirichlet Process," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1131–1154, Sept. 2008.

[49] Amr Ahmed, Liangjie Hong, and Alexander J. Smola, "Nested Chinese Restaurant Franchise Process: Applications to User Tracking and Document Modeling," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013, pp. 1426–1434.

[50] John William Paisley, Chong Wang, David M. Blei, and Michael I. Jordan, "Nested Hierarchical Dirichlet Processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 256–270, 2015.

[51] Tengfei Ma, Issei Sato, and Hiroshi Nakagawa, "The Hybrid Nested/Hierarchical Dirichlet Process and its Application to Topic Modeling with Word Differentiation.," in *AAAI*, 2015, pp. 2835–2841.

[52] Griffin and Steel, "Order-Based Dependent Dirichlet Processes," Tech. Rep., 2006.

[53] Lu Ren, David B. Dunson, and Lawrence Carin, "The dynamic hierarchical Dirichlet process," in *Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, 2008, pp. 824–831.

[54] Richard Socher, Andrew L. Maas, and Christopher D. Manning, "Spectral Chinese Restaurant Processes: Nonparametric Clustering Based on Similarities," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, 2011, pp. 698–706.

[55] Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei, "Spatial distance dependent Chinese restaurant processes for image segmentation," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. 2011, pp. 1476–1484, Curran Associates, Inc.

[56] Amr Ahmed and Eric P. Xing, "Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: With Applications to Evolutionary Clustering," in *Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA*, 2008, pp. 219–230.

[57] Romain Thibaux and Michael I. Jordan, "Hierarchical Beta Processes and the Indian Buffet Process," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, 2007, pp. 564–571.

[58] John Paisley and Lawrence Carin, "Nonparametric Factor Analysis with Beta Process Priors," in *Proceedings of the 26th International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 777–784, ACM.

[59] Mingyuan Zhou, Haojun Chen, Lu Ren, Guillermo Sapiro, Lawrence Carin, and John W. Paisley, "Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, Curran Associates, Inc.

[60] Mingyuan Zhou, Hongxia Yang, Guillermo Sapiro, David B. Dunson, and Lawrence Carin, "Dependent Hierarchical Beta Process for Image Interpolation and Denoising," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, 2011, pp. 883–891.

[61] Sunil Gupta, Dinh Phung, and Svetha Venkatesh, "Factorial Multi-Task Learning : A Bayesian Nonparametric Approach," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013, pp. 657–665.

[62] Zoubin Ghahramani and Thomas L. Griffiths, "Infinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schlkopf, and J. C. Platt, Eds. 2005, pp. 475–482, MIT Press.

[63] Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis, "Modeling Dyadic Data with Binary Latent Factors," in *Advances in Neural Information Processing Systems 19*, B. Schlkopf, J. C. Platt, and T. Hoffman, Eds. 2007, pp. 977–984, MIT Press.

[64] Kurt Miller, Michael I. Jordan, and Thomas L. Griffiths, "Nonparametric Latent Feature Models for Link Prediction," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, pp. 1276–1284, Curran Associates, Inc.

[65] Piyush Rai and Hal Daume, "Multi-Label Prediction via Sparse Infinite CCA," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, pp. 1518–1526, Curran Associates, Inc.

[66] Frank Wood, Thomas L. Griffiths, and Zoubin Ghahramani, "A Non-Parametric Bayesian Method for Inferring Hidden Causes," in *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*, 2006.

[67] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[68] Lu Ren, Yingjian Wang, Lawrence Carin, and David B. Dunson, "The Kernel Beta Process," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. 2011, pp. 963–971, Curran Associates, Inc.

[69] Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan, "The Phylogenetic Indian Buffet Process: A Non-exchangeable Nonparametric Prior for Latent Features," in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States, 2008, UAI'08, pp. 403–410, AUAI Press.

[70] S.J. Gershman, P.I. Frazier, and D.M. Blei, "Distance Dependent Infinite Latent Feature Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 334–345, Feb. 2015.

[71] Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani, "Dependent Indian Buffet Processes," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, 2010, pp. 924–931.

[72] Daryl J Daley and David Vere-Jones, *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*, Springer Science & Business Media, 2007.

[73] Peter J. Diggle, Thomas Fiksel, Pavel Grabarnik, Yosihiko Ogata, Dietrich Stoyan, and Masaharu Tanemura, "On Parameter Estimation for Pairwise Interaction Point Processes," *International Statistical Review / Revue Internationale de Statistique*, vol. 62, no. 1, pp. 99, Apr. 1994.

[74] Brian D Ripley, "Modelling spatial patterns," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 172–212, 1977.

[75] Alan G Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[76] Alan G. Hawkes and David Oakes, "A Cluster Process Representation of a Self-Exciting Process," *Journal of Applied Probability*, vol. 11, no. 3, pp. 493–503, 1974.

[77] Yosihiko Ogata, "Seismicity analysis through point-process modeling: A review," in *Seismicity Patterns, Their Statistical Significance and Physical Meaning*, pp. 471–507. Springer, 1999.

[78] JFC Kingman, "On doubly stochastic Poisson processes," in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1964, vol. 60, pp. 923–930, Cambridge Univ Press.

[79] Jerzy Neyman and Elizabeth L Scott, "Statistical approach to problems of cosmology," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–43, 1958.

[80] RP Adams, *Kernel Methods for Nonparametric Bayesian Inference of Probability Densities and Point Processes*, Ph.D. thesis, University of Cambridge, 2009.

[81] Jasper Snoek, Richard Zemel, and Ryan P Adams, "A determinantal point process latent variable model for inhibition in neural spiking data," in *NIPS*, 2013, pp. 1932–1940.

[82] Ke Zhou, Hongyuan Zha, and Le Song, "Learning Triggering Kernels for Multi-dimensional Hawkes Processes," in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1301–1309.

[83] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[84] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[85] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.* 1986, vol. 11, pp. 49–52, IEEE.

[86] Christopher D. Manning and Hinrich Schtze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, USA, 1999.

[87] Camilla Landn, "Bond pricing in a hidden Markov model of the short rate," *Finance and Stochastics*, vol. 4, no. 4, pp. 371–389, Aug. 2000.

[88] M. R. Hassan and B. Nath, "Stock market forecasting using hidden Markov model: A new approach," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, Sept. 2005, pp. 192–196.

[89] A. M. Gonzalez, A. M. S. Roque, and J. Garcia-Gonzalez, "Modeling and forecasting electricity prices with input/output hidden Markov models," *IEEE Transactions on Power Systems*, vol. 20, no. 1, pp. 13–24, Feb. 2005.

[90] Stuart Geman and Donald Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[91] Balvant Rajani and Yehuda Kleiner, "Comprehensive review of structural deterioration of water mains: Physically based models," *Urban water*, vol. 3, no. 3, pp. 151–164, 2001.

[92] U. Shamir and C.D.D. Howard, "An analytic approach to scheduling pipe replacement," 1979, pp. 71(5),248–258, Journal of AWWA.

[93] K. Mavin, "Predicting the failure performance of individual water mains," *Urban Water Research Association of Australia*, , no. 114, 1996.

[94] AJ Kettler and IC Goulter, "An analysis of pipe breakage in urban water distribution networks," *Canadian Journal of Civil Engineering*, vol. 12, no. 2, pp. 286–293, 1985.

[95] A. G. Constantine, "Pipeline reliability: Stochastic models in engineering technology and management," Singapore: World Scientific, 1996.

[96] Rui Wang, Weishan Dong, Yu Wang, Ke Tang, and Xin Yao, "Pipe failure prediction: A data mining method," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference On*. 2013, pp. 1208–1218, IEEE.

[97] David J Aldous, *Exchangeability and Related Topics*, Springer, 1985.

[98] Shuang-Hong Yang and Hongyuan Zha, "Mixture of Mutually Exciting Processes for Viral Diffusion.," *ICML*, vol. 28, pp. 1–9, 2013.

[99] Patrick Hewlett, "Clustering of order arrivals, price impact and trade path optimisation," in *Workshop on Financial Modeling with Jump Processes, Ecole Polytechnique*, 2006, pp. 6–8.

[100] Honglin Yu, Lexing Xie, and Scott Sanner, "The Lifecyle of a Youtube Video: Phases, Content and Popularity," in *AAAI Conference on Web and Social Media*, 2015.

[101] D Vere-Jones, "An Introduction to the Theory of Point Processes," *Springer Ser. Statist., Springer, New York*, 1988.

[102] P. Orbanz and D.M. Roy, "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 437–461, Feb. 2015.

[103] Sergio Bacallado, Stefano Favaro, Lorenzo Trippa, and others, "Bayesian nonparametric analysis of reversible Markov chains," *The Annals of Statistics*, vol. 41, no. 2, pp. 870–896, 2013.

[104] Persi Diaconis and David Freedman, "De Finetti's theorem for Markov chains," *The Annals of Probability*, pp. 115–130, 1980.

[105] Sandy L Zabell, "Characterizing Markov exchangeable sequences," *Journal of Theoretical Probability*, vol. 8, no. 1, pp. 175–178, 1995.

[106] David M. Blei and Peter I. Frazier, "Distance Dependent Chinese Restaurant Processes," *J. Mach. Learn. Res.*, vol. 12, pp. 2461–2488, Nov. 2011.

[107] Y. Ogata, "On Lewis' simulation method for point processes," *Information Theory, IEEE Transactions on*, vol. 27, pp. 23–31, Jan. 1981.

[108] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda, "Learning Systems of Concepts with an Infinite Relational Model," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, Boston, Massachusetts, 2006, AAAI'06, pp. 381–388, AAAI Press.

[109] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei, "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes.," in *NIPS*, 2004.

[110] Junchi Yan, Yu Wang, Ke Zhou, Jin Huang, Chunhua Tian, Hongyuan Zha, and Weishan Dong, "Towards effective prioritizing water pipe replacement and rehabilitation," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. 2013, pp. 2931–2937, AAAI Press.

[111] Yehuda Kleiner and Balvant Rajani, "Comprehensive review of structural deterioration of water mains: Statistical models," *Urban water*, vol. 3, no. 3, pp. 131–150, 2001.

[112] Bin Li, Bang Zhang, Zhidong Li, Yang Wang, Fang Chen, and Dammika Vitanage, "PRIORITISING WATER PIPES FOR CONDITION ASSESSMENT WITH DATA ANALYTICS," *OzWater*, 2015.

[113] Peter J Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[114] Adrian J Baddeley and MNM Van Lieshout, "Area-interaction point processes," *Annals of the Institute of Statistical Mathematics*, vol. 47, no. 4, pp. 601–619, 1995.

[115] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS computational biology*, vol. 7, no. 3, pp. e1001110, 2011.

[116] Jyri J Kivinen, Erik B Sudderth, and Michael I Jordan, "Learning multiscale representations of natural scenes using Dirichlet processes," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference On*. 2007, pp. 1–8, IEEE.

[117] Derek Hao Hu, Xian-Xing Zhang, Jie Yin, Vincent Wenchen Zheng, and Qiang Yang, "Abnormal Activity Recognition Based on HDP-HMM Models.," in *IJCAI*, 2009, pp. 1715–1720.

[118] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky, "A sticky HDP-HMM with application to speaker diarization," *The Annals of Applied Statistics*, pp. 1020–1056, 2011.

[119] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, "An HDP-HMM for systems with state persistence," in *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 312–319, ACM.

[120] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani, "Beam sampling for the infinite hidden Markov model," in *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 1088–1095, ACM.

[121] Liangda Li and Hongyuan Zha, "Energy Usage Behavior Modeling in Energy Disaggregation via Marked Hawkes Process," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[122] J. Zico Kolter and Matthew J. Johnson, "REDD: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*. 2011, vol. 25, pp. 59–62, Citeseer.

[123] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[124] David M Blei, Michael I Jordan, and others, "Variational inference for Dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[125] Yee Whye Teh, Kenichi Kurihara, and Max Welling, "Collapsed Variational Inference for HDP.," in *NIPS*, 2007, pp. 1481–1488.

[126] Emmanuel Bacry, Stphane Gaffas, Iacopo Mastromatteo, and Jean-Franois Muzy, "Mean-field inference of Hawkes point processes," *Journal of Physics A: Mathematical and Theoretical*, vol. 49, no. 17, pp. 174006, 2016.

[127] Thomas L. Griffiths and Zoubin Ghahramani, "The Indian Buffet Process: An Introduction and Review," *J. Mach. Learn. Res.*, vol. 12, pp. 1185–1224, July 2011.

[128] Jurgen V. Gael, Yee W. Teh, and Zoubin Ghahramani, "The Infinite Factorial Hidden Markov Model," in *Advances in Neural Information Processing*

Systems 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. 2009, pp. 1697–1704, Curran Associates, Inc.

[129] Emily B. Fox, Michael I. Jordan, Erik B. Sudderth, and Alan S. Willsky, "Sharing Features among Dynamical Systems with Beta Processes," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, pp. 549–557, Curran Associates, Inc.