TECHNISCHE
UNIVERSITÄT
DARMSTADT

NONPARAMETRIC BAYESIAN LATENT FACTOR MODELS FOR NETWORK RECONSTRUCTION

Dem Fachbereich Elektrotechnik und Informationstechnik der
TECHNISCHE UNIVERSITÄT DARMSTADT

zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
vorgelegte Dissertation

von

SIKUN YANG, M.PHIL.

Referent: Prof. Dr. techn. Heinz Köppl
Korreferent: Prof. Dr. Kristian Kersting

Darmstadt 2019

The work of Sikun Yang was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 668858 at the Technische Universsität Darmstadt, Germany.

Yang, Sikun : *Nonparametric Bayesian Latent Factor Models for Network Reconstruction*,
Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2019
URN: `urn:nbn:de:tuda-tuprints-96957`
Tag der mündlichen Prüfung: 11.12.2019

Veröffentlicht unter CC BY-SA 4.0 International
`https://creativecommons.org/licenses/by-sa/4.0/`

ABSTRACT

This thesis is concerned with the statistical learning of probabilistic models for graph-structured data. It addresses both the theoretical aspects of network modelling–like the learning of appropriate representations for networks–and the practical difficulties in developing the algorithms to perform inference for the proposed models.

The first part of the thesis addresses the problem of discrete-time dynamic network modeling. The objective is to learn the common structure and the underlying interaction dynamics among the entities involved in the observed temporal network. Two probabilistic modeling frameworks are developed. First, a Bayesian nonparametric framework is proposed to capture the *static* latent community structure and the evolving node-community memberships over time. More specifically, the hierarchical gamma process is utilized to capture the underlying intra-community and inter-community interactions. The appropriate number of latent communities can be automatically estimated via the inherent shrinkage mechanism of the hierarchical gamma process prior. The gamma Markov process are constructed to capture the evolving node-community relations. As the Bernoulli-Poisson link function is used to map the *binary* edges to the latent parameter space, the proposed method scales with the number of non-zero edges. Hence, the proposed method is particularly well-fitted to model large sparse networks. Moreover, a time-dependent hierarchical gamma process dynamic network model is proposed to capture the birth and death dynamics of the underlying communities. For performance evaluation, the proposed methods are compared with state-of-the-art statistical network models on both synthetic and real-world data.

In the second part of the thesis, the main objective is to analyze continuous-time event-based dynamic networks. A fundamental problem in modeling such continuously-generated temporal interaction events data is to capture the *reciprocal* nature of the interactions among entities–the actions performed by one individual toward another increase the probability that an action of the same type to be returned. Hence, the mutually-exciting Hawkes process is utilized to capture the reciprocity between each pair of individuals involved in the observed dynamic network. In particular, the base rate of the Hawkes process is built upon the latent parameters inferred using the hierarchical gamma process edge partition model, to capture the underlying community structure. Moreover, each interaction event between two individuals is augmented with a pair of latent variables, which will be referred to as latent patterns, to indicate which of their involved communities lead to the occurring of that interaction. Accordingly, the proposed model allows the excitatory effects of each interaction on its opposite direction are determined by its latent patterns. Efficient Gibbs sampling and Expectation Maximization algorithms are developed to perform inference. Finally, the evaluations performed on the real-world data demonstrate the interpretability and competitive performance of the model compared with state-of-the-art methods.

In the third part of this thesis, the objective is to analyze the common structure of multiple related data sources under the generative framework. First, a Bayesian nonparametric group factor analysis method is developed to factorize multiple related groups of data into the common latent factor space. The hierarchical beta Bernoulli process is exploited to induce sparsity over the group-specific factor loadings to strengthen the model interpretability. A collapsed variational inference scheme is proposed to perform efficient inference for large-scale data analysis in real-world applications. Moreover, a Poisson gamma memberships framework is investigated for joint modelling of network and related node features.

ZUSAMMENFASSUNG

Die Dissertation beschäftigt sich mit dem statistischen Lernen von Wahrscheinlichkeitsmodellen für graphisch strukturierte Daten. Es befasst sich sowohl mit den theoretischen Aspekten der Netzwerkmodellierung - wie dem Erlernen geeigneter Darstellungen für Netzwerke - als auch mit den praktischen Schwierigkeiten bei der Entwicklung der Algorithmen zur Durchführung von Inferenzen für die vorgeschlagenen Modelle.

Der erste Teil die Dissertation befasst sich mit dem Problem der zeitdiskreten dynamischen Netzwerkmodellierung. Ziel ist es, die gemeinsame Struktur und die zugrunde liegende Dynamik der am beobachteten zeitlichen Netzwerk beteiligten Entitäten zu lernen. Es werden zwei probabilistische Modellierungsrahmen entwickelt. Zunächst wird ein Bayes'sches nichtparametrisches Framework vorgeschlagen, um die *statische* latente Community-Struktur und die sich im Laufe der Zeit entwickelnden Node-Community-Mitgliedschaften zu erfassen. Insbesondere wird der hierarchische Gamma-Prozess verwendet, um die zugrunde liegenden innergemeinschaftlichen und zwischengemeinschaftlichen Interaktionen zu erfassen. Die geeignete Anzahl latenter Gemeinschaften kann über den inhärenten Schrumpfungsmechanismus des hierarchischen Gamma-Prozesses vor geschätzt werden. Der Gamma-Markov-Prozess ist so aufgebaut, dass er die sich entwickelnden Knoten-Community-Beziehungen erfasst. Da die Bernoulli-Poisson-Beziehung verwendet wird, um die binären Kanten in den latenten Parameterraum abzubilden, skaliert die vorgeschlagene Methodik mit der Anzahl der Kanten. Daher ist die vorgeschlagene Methodik gut geeignet, um große dünnbesetz Netzwerke zu modellieren. Darüber hinaus wird ein zeitabhängiges dynamisches Netzwerkmodell für hierarchische Gamma-Prozesse vorgeschlagen, um die Geburts- und Todesdynamik der zugrunde liegenden Gemeinschaften zu erfassen. Zur Leistungsbewertung werden die vorgeschlagenen Methoden mit den neuesten statistischen Netzwerkmodellen für synthetische und reale Daten verglichen.

Im zweiten Teil die Dissertation geht es vor allem darum, zeitkontinuierliche ereignisbasierte dynamische Netzwerke zu analysieren. Ein grundlegendes Problem bei der Modellierung solcher kontinuierlich erzeugten zeitlichen Interaktionsereignisse besteht darin, die reziproke Art der Wechselwirkung Interaktionen zwischen Entitäten zu erfassen. Der sich gegenseitig erregende Hawkes-Prozess wird verwendet, um die Reziprozität zwischen jedem Paar von Personen in dem beobachteten dynamischen Netzwerk zu erfassen. Insbesondere basiert der Hawkes-Prozess auf den latenten Parametern, die unter Verwendung des hierarchischen Gamma-Prozess-Kantenpartitionsmodells abgeleitet wurden, um die zugrunde liegende Community-Struktur zu erfassen.

Darüber hinaus wird jedes Ereignis zwischen zwei Individuen mit einem Paar aus latenten Variablen versehen, welche als latente Muster zu verstehen sind. Das vorgeschlagene Modell ermöglicht, dass die anregenden Effekte jedes Ereignisses durch seine latenten Muster bestimmt werden. Effiziente Gibbs-Abtast- und Erwartungswert-Maximierungs-Algorithmen werden entwickelt, um Inferenzen durchzuführen. Schließlich belegen die Auswertungen der realen Daten die hohe Wettbewerbsfähigkeit und eine Leistung auf dem neuesten Stand der Technik.

Der dritte Teil die Dissertation stellt sich das Ziel, die gemeinsame Struktur von multiplen verwandtden Datenquellen unter einem generativen Rahmen zu analysieren. Zunächst wird ein Bayes'sches Verfahren zur Analyse nichtparametrischer Gruppenfaktoren entwickelt, um mehrere verwandte Datengruppen in den gemeinsamen Latenzfaktorraum zu zerlegen. Der hierarchische

Beta-Bernoulli-Prozess wird ausgenutzt, um die Dünnbesetztheit gegenüber dem gruppenspezifischen Faktor zu induzieren. Es wird ein reduziertes Variations-Inferenz-Schema vorgeschlagen, um eine effiziente Inferenz für eine Datenanalyse in großem Maßstab in realen Anwendungen durchzuführen. Darüber hinaus untersuchen wir ein Poisson-Gamma-Mitgliedschafts-Framework für die gemeinsame Modellierung von Netzwerk- und verwandten Knotenmerkmalen.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge Prof. Heinz Koeppl for giving me the opportunity to pursue my research interests. I would like to thank him for his countless encouragement, support, and advices.

I am grateful to Prof. Martha Gibson for her generous and constant support for my research.

I am sincerely thankful to Markus Baier and Christine Cramer for their generous helping with administrative and technical matters. I would like to thank all the former and current members of the BCS lab, Christiane Hübner, Bastian Alt, Leo Bronstein, Jascha Diemer, Maleen Hanst, Nikita Kruk, Francois-Xavier Lehr, Dominik Linzner, Tim Prangemeier, Adrian Šošić, Christian Wildner, Mark Sinzger, Hameer Abbasi, Volodymir Volchenko, Tabea Treppmann, Sara Al-Sayed, Nurgazy Sulaimanov, Derya Altintan, Megan Bailey, Nicolai Engelmann, Kilian Heck, Lukas Köhs, Felix Reinhardt, Markus Röder, Michael Schmidt, Rogier Schoeman, Anam Tahir, Klaus-Dieter Voss, Özdemir Cetin, Yuri Maistrenko, Sunil Kumar. Specifically, I am sincerely thankful to Adrian Šošić and Nikita Kruk for their generous helping with my research projects, and to Nurgazy Sulaimanov for his countless encouragement.

I would also like to thank all my friends, Shuanglai Fu, Xiaolan Ross, Prof. Hong-Yang Chao, Prof. Mingyuan Zhou, Zhitang Chen, Prof. Nenggan Zheng, Yang Han, Yunhai Wu, Ruo Yi, Junbo Lu.

Last, but by no means least, I would like to thank my family for their love and constant support.

CONTENTS

1 INTRODUCTION 1
- 1.1 Thesis overview and contributions . . . 2
- 1.2 Publications . . . 4

2 BACKGROUND AND RELATED WORK 5
- 2.1 Fundamental definitions . . . 5
- 2.2 Probability Distributions, Stochastic Processes and Temporal Point Processes . . . 5
- 2.3 Network Models . . . 11

3 DISCRETE TIME DYNAMIC NETWORK MODELS 15
- 3.1 Related work . . . 16
- 3.2 The Dynamic Poisson Gamma Memberships Model . . . 16
- 3.3 The Dependent Relational Gamma Process Model . . . 24
- 3.4 Experiments . . . 30
- 3.5 Conclusions . . . 42

4 CONTINUOUS TIME DYNAMIC NETWORK MODELS 47
- 4.1 Related work . . . 48
- 4.2 Hawkes Processes with the Hierarchical Gamma Process Edge Partition Model . . 49
- 4.3 The Hawkes Edge Partition Model . . . 50
- 4.4 Inference . . . 52
- 4.5 Experiments . . . 57
- 4.6 Conclusions . . . 59

5 STOCHASTIC GRADIENT MARKOV CHAIN MONTE CARLO FOR DISCRETE TIME NETWORK MODELS 63
- 5.1 Stochastic Gradient Markov Chain Monte Carlo . . . 63
- 5.2 The Dynamic Dirichlet Edge Paritition Model . . . 64
- 5.3 Inference . . . 65
- 5.4 Experiments . . . 70
- 5.5 Conclusions . . . 72

6 NONPARAMETRIC BAYESIAN GROUP FACTOR ANALYSIS 75
- 6.1 Nonparametric Bayesian Group Factor Analysis . . . 76
- 6.2 Collapsed Variational Inference . . . 77
- 6.3 Experiments . . . 85
- 6.4 Conclusions . . . 93

7 JOINT MODELS FOR NETWORK EDGES AND NODE FEATURES 95
- 7.1 The Poisson Gamma Memberships Model for Network Edges and Node Features . 95
- 7.2 Inference . . . 98
- 7.3 Experiments . . . 99
- 7.4 Conclusions . . . 99

8 CONCLUSIONS AND FUTURE RESEARCH 101

ACRONYMS 103

BIBLIOGRAPHY 107

CURRICULUM VITÆ 115

ERKLÄRUNG LAUT §9 PROMOTIONSORDNUNG 117

1

INTRODUCTION

There has been considerable interest in network analysis because many complicated physical, social and biological phenomena, such as protein-protein interactions and friendship relations among individuals, can be represented as networks. A network is composed of nodes and edges between them. To date, a large amount of work has been done on the analysis of static networks, which either represent an aggregated view of networks for a time period, or a single network snapshot observed at a time point. As internet and biological technologies advance, a rich collection of graph-structured data has become available for modelling and understanding network formation and evolving processes.

On the one aspect, instead of observing a single aggregated view, a time-varying network either consists of a set of snapshots (discrete-time networks) collected at multiple time points, or represents a continuous-time evolving network, in which each edge is associated with a timestamp. On the other aspect, auxiliary node features, such as user profiles in online social networks or gene-expression data along with gene regulatory networks, are also available to be leveraged into network modelling when the observed network is incomplete.

The main objectives of dynamic network models are to estimate the common structure, while at the same time to capture the underlying dynamic interactions among nodes. In addition, real-world networks are often extremely sparse (only a small fraction of network entries are non-zeros), and typically exhibit high degree (number of edges per node) heterogeneity–some nodes have a large number of connections, while most nodes have very few edges. Therefore, it is highly desirable to develop methods that not only can capture the evolving node behaviour and interpret the edge formation mechanisms, but also truly scale to large sparse networks.

For discrete-time networks, prior works include deterministic approaches, such as exponential random graph models (Guo et al. 2007), matrix and tensor factorization based methods (Dunlavy et al. 2011), and statistical models (Sarkar et al. 2006; Xing et al. 2010; J. R. Foulds et al. 2011; Heaukulani et al. 2013). Among these methods, statistical network models received an amount of attention because these models have great flexibility and show favorable interpretability by providing uncertainty estimates for the discovered latent parameters. Moreover, these models based upon generative mechanisms often perform well in predicting missing edges and forecasting future unseen snapshots. The statistical models for discrete-time networks mainly include class-based and feature-based models. The class-based models, such as the dynamic stochastic block model (dSBM) (Xing et al. 2010), assign the nodes of a network into a finite number of classes, and determine the edge between each pair of two nodes entirely by their assigned classes. The dSBM captures evolving node-class assignments using state space models. The feature-based models, such as the dynamic latent feature relational model (dLFRM) (Miller et al. 2009), represent each node with a binary vector, which naturally captures the underlying overlapping community structure because each node can belong to multiple communities (features). In the dLFRM, the node-community memberships independently evolve over time according to the factorial hidden Markov model (Zoubin Ghahramani et al. 1996). Despite having expressive representations, the dynamic feature relational models map the binary edges to the latent space using the logistic link function that scales quadratically with the number of nodes. Hence, it is unrealistic to apply dLFRMs in analysis of sparse networks with very large number of nodes. In addition, the dLFRM can only characterize the binary node-community memberships, and thus fail to capture the degrees of node-memberships to multiple communities. For instance, an individual is more likely to connect to another individual if they both have high degrees in a community like "rock music" but low degrees in a community like "classical music", than an individual with low degree in "rock music" but high degree in "classical music". Therefore, the dLFRMs cannot capture such differences in the degrees of node-community memberships because the three individuals have the same binary feature representations. Furthermore, the node-community memberships are expected to change smoothly when the corresponding nodes join or leave the communities. The dLFRMs cannot capture such smooth evolving behaviour without modelling the degrees of node memberships appropriately. In addition, discrete-time network models aggregate timestamped relational events to form network snapshots, which unavoidably discard a significant amount of information. For example, if Bob emails Alice, then Alice is more likely to send an email to Bob in the near future (reciprocity). To capture such reciprocating nature of temporal interactions, it makes more sense to model a collection of temporal interaction events that implicitly form a continuous-time network, rather than discrete snapshots collected at regular time intervals.

In this thesis, the main contribution is to develop statistical models, in which each entity is endowed with an expressive node-community memberships vector. Consequently, the proposed model not only captures the underlying overlapping community structure, but also measures the degrees of each node's memberships to the multiple communities. Accordingly, the proposed model allows each node-community membership to evolve smoothly over time. In addition, the proposed model can effectively leverage the sparsity manifested in large networks, and thus admits a simple-yet-efficient Gibbs sampling algorithm to perform inference. Moreover, this thesis also presents models for continuous-time event-based networks. To this end, a temporal point process-based model is proposed to capture the underlying community structure behind temporal interactions. In the proposed model, each event between a pair of individuals is either driven by their respective affiliated communities or triggered by the past opposite interactions between them. Another challenging task in network modelling is to reconstruct incomplete network edges using available node-specific side information, such as node features. For example, in computational system biology, interacting proteins tend to be linked to similar phenotypes and participating in similar functions. Hence, to predict protein-protein interaction (PPI) networks using available protein sequence data and structural information is a difficult problem that scientists strive to address. In this thesis, a probabilistic generative model is developed to jointly model the network data with node features.

1.1 THESIS OVERVIEW AND CONTRIBUTIONS

The goal of this thesis is to develop statistical models such that:

1. Development of discrete-time network models that capture the underlying community structure and the node evolving behaviours, and also characterize the edge formation.

2. Development of continuous-time network models that characterize the latent community structure behind interactions among nodes, and capture the reciprocating interactions between each pair of nodes.

3. Development of models that jointly capture the generative process of a network and its associated node-specific side information.

4. Development of scalable inference schemes for the developed dynamic network models.

5. Development of models for joint modelling of multiple related groups of data.

This thesis will develop statistical network models that fulfil the above objectives. We start by reviewing the related work. The developed statistical models are presented in Chapter 3 to 7. The following subsections describe the chapters for each model in more detail.

1.1.1 *Models for discrete-time networks (Chapter 3)*

In this chapter, a statistical model for discrete-time networks is presented. The proposed model represents each node by a nonnegative node-community memberships vector that enables us to capture the underlying overlapping community structure and also measures the degrees of node memberships. The proposed model characterizes both intra-community and inter-community interactions using a positive weight matrix that builds upon the hierarchical gamma process. Hence, this model allows any two nodes that have no common affiliated communities to connect through the inter-community interactions. By the intrinsic shrinkage mechanism of the hierarchical gamma process, the proposed model can automatically infer an appropriate number of latent communities in a Bayesian nonparametric way. Using the Bernoulli-Poisson link (BPL) function, the model maps the binary network edges to the latent relational space. As the inference only needs to be performed for non-zero network entries, the BPL function makes the proposed model appealing for modelling large sparse networks. The proposed model captures the smooth evolving behaviour of each node-community membership using a gamma Markov process. Although the exact inference for the proposed model is analytically intractable, a simple-yet-efficient Gibbs sampling scheme with full local conjugacy using the data augmentation and marginalization strategies, is developed to perform inference. Moreover, the proposed model can be readily generalized to count-valued and to positive real-valued networks using the Poisson randomized gamma function.

1.1.2 *Models for continuous-time networks (Chapter 4)*

In this chapter, a temporal point process-based model is derived for continuous-time event-based dynamic networks. In particular, this model captures the underlying community structure behind temporal interactions among nodes by incorporating such structural information into the base intensity of the temporal point process model. Then, this model captures the reciprocating interactions between each pair of two nodes using a mutually-exciting Hawkes process. The proposed model assumes that each event between two nodes is either driven by these two nodes' affiliated communities or triggered by the opposite past interactions of the same type. Therefore, the proposed model can flexibly allow the interaction dynamics between two nodes to be modulated by their affiliated communities. In addition, our model can also incorporate the available node features or node-generated contents via the prior specification. Both efficient Gibbs sampling and Expectation-Maximization inference schemes are developed to perform inference.

1.1.3 *Stochastic gradient MCMC inference for dynamic network models (Chapter 5)*

In this chapter, we extend the bilinear Poisson factorization model for dynamic networks by constructing node-community memberships via the Dirichlet Markov chain structure. Moreover, the hierarchical beta gamma prior is utilized to prevent the over estimation of the number of latent communities. This novel framework enables us to derive fast stochastic gradient Markov chain Monte Carlo algorithms using the expanded-mean and the reduced-mean reparameterization strategies.

1.1.4 *Models for joint modelling of multiple related groups of data (Chapter 6)*

In this chapter, a Bayesian nonparametric model for joint modelling of multiple related groups of data is presented. The proposed model captures both the group-specific signals and the underlying common structure of multiple related grouped data under the group factor analysis framework. Using the hierarchical beta-Bernoulli process prior, this framework can automatically infer the number of latent factors by data itself, and effectively induce the sparsity over the factor loadings to improve the interpretability. For large-scale group factor analysis, a collapsed variational inference scheme is developed to perform inference in the proposed framework.

1.1.5 *Models for networks and node features (Chapter 7)*

In this chapter, we extend the bilinear Poisson factorization model for static networks to jointly model a static network and its associated node features, which are considered to be observed here. In the proposed hierarchical Bayesian model, both the observed network and node features are factorized in the common latent relational space by representing each node with a positive node-community memberships vector. This model enables us to reconstruct missing edges in network data using available node features. Moreover, the proposed model can be straightforwardly extended to model dynamic or multi-relational networks.

1.2 PUBLICATIONS

The following articles have been written during the course of the author's doctoral studies.

1. Sikun Yang and Heinz Koeppl (2018b). "A Poisson Gamma Probabilistic Model for Latent Node-Group Memberships in Dynamic Networks". In: *AAAI Conference on Artificial Intelligence*, pp. 4366–4373.

2. Sikun Yang and Heinz Koeppl (2018c). "Dependent Relational Gamma Process Models for Longitudinal Networks". In: *International Conference on Machine Learning (ICML)*, pp. 5551–5560.

3. S. Yang and H. Koeppl (2018a). "Collapsed Variational Inference for Nonparametric Bayesian Group Factor Analysis". In: *IEEE International Conference on Data Mining (ICDM)*, pp. 687–696.

4. Sikun Yang and Heinz Koeppl (2019b). "The Hawkes Edge Partition Model for Continuous-time Event-based Temporal Networks". In: *Submitted.*

5. Sikun Yang and Heinz Koeppl (2019a). "An Empirical Study of Stochastic Gradient MCMC Algorithms for the Dynamic Edge Partition Models". In: *Submitted.*

2

BACKGROUND AND RELATED WORK

In this chapter, we first introduce the fundamental definitions and terminology that will be used in this thesis. Then, we survey previous work for modelling static networks, discrete-time networks, continuous-time event-based networks and network models that incorporate node features.

2.1 FUNDAMENTAL DEFINITIONS

2.1.1 *Static networks*

Formally, a static network can be represented by a graph $G \equiv (\mathcal{V}, \mathcal{E})$ that is composed of a set of nodes $\mathcal{V}$ and the edges $\mathcal{E}$ between these nodes. Here, $V \equiv |\mathcal{V}|$ denotes the number of nodes, and $E \equiv |\mathcal{E}|$ the number of edges. A graph G can be represented by an adjacency matrix $A \in \{0, 1\}^{V \times V}$, where $A_{uv} = 1$ if an edge is observed between nodes u and v, and $A_{uv} = 0$ otherwise. A graph is undirected if and only if $A_{uv} = A_{vu}$, and directed otherwise. For undirected graphs, the degree of a node is defined by the number of nodes that connect with this given node. In directed graphs, we define the in-degree of a node by the number of nodes that have edges incoming to this given node, and the out-degree of a node by the number of nodes outgoing from this given node. We use the term node, vertex and entity interchangeably. Similarly, we use link and edge interchangeably, and also use group and community interchangeably.

2.1.2 *Discrete-time networks*

A discrete-time network is represented as a sequence of binary adjacency matrices $\{A^{(t)}\}_{t=1}^{T}$, where $A^{(t)} \in \{0, 1\}^{V \times V}$ for each time point $t = 1, \ldots, T$. The adjacency matrix $A^{(t)}$ has entries $A_{uv}^{(t)} = 1$ if an edge from nodes u to v is present at time t, and $A_{uv}^{(t)} = 0$ otherwise.

2.1.3 *Continuous-time event-based networks*

A dynamic network evolving continuously over time, such as email communication networks and exchange of messages on online social networks, can be observed through a sequence of interaction events between pairs of nodes at recorded timestamps. For instance, a set of temporal interaction events can be represented as $\{(t_i, s_i, d_i)\}_{i=1}^{N}$, where N is the number of events, and (t_i, s_i, d_i) denotes an event directed from node s_i (sender) to node d_i (receiver) at timestamp t_i.

2.2 PROBABILITY DISTRIBUTIONS, STOCHASTIC PROCESSES AND TEMPORAL POINT PROCESSES

In this section, we describe the distributions, the nonparmetric Bayesian priors, and the data augmentation and marginalization strategies covered in this thesis. When expressing the full conditionals for Gibbs sampling we will use the shorthand "–" to denote all other variables. We use "·" as a index summation shorthand, e.g., $x_{\cdot j} = \sum_i x_{ij}$.

2.2.1 *The Poisson Distribution*

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if for $k = 0, 1, 2, \dots,$ the probability mass function of X is defined by

$$f_X(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{2.2.1}$$

where $x!$ is the factorial of x. The mean and variance of X are $\mathsf{E}[X] = \lambda$, and $\mathsf{Var}[X] = \lambda$, respectively.

2.2.2 *The Gamma Distribution*

A random variable X drawn from a gamma distribution with shape parameter a and scale parameter $1/b$ is denoted as $X \sim \text{Gamma}(a, 1/b)$, and has a probability density function as

$$f_X(x \mid a, 1/b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \tag{2.2.2}$$

where $\Gamma(\cdot)$ denotes the Gamma function. The mean and variance of X are $\mathsf{E}[X] = a/b$ and $\mathsf{Var}[X] = a/b^2$, respectively.

2.2.3 *The Dirichlet distribution*

A K–dimensional random vector, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, where $\pi_i > 0$ and $\sum_{i=1}^K \pi_i = 1$, is said to take values in $(K-1)$–dimensional–simplex. The Dirichlet distribution of dimension K is a continuous probability distribution on $(K-1)$–dimensional simplex, and has a probability density function as

$$f(\boldsymbol{\pi} \mid \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}. \tag{2.2.3}$$

Let $\alpha_0 = \sum_k \alpha_k$. The mean and variance of an element of $\boldsymbol{\pi}$ are $\mathsf{E}[\pi_k] = \frac{\alpha_k}{\alpha_0}$ and $\mathsf{Var}[\pi_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$, respectively.

2.2.4 *The Negative-Binomial distribution*

A negative-binomial distributed random variable $X \sim \text{NB}(r, p)$ has a probability mass function as

$$f_X(x|r, p) = \frac{\Gamma(x + r)}{x!\Gamma(r)} p^x (1-p)^r, \tag{2.2.4}$$

where $x \in \mathbb{Z}_{\geq 0}$. The mean and variance of X are $\mathsf{E}[X] = rp/(1-p)$ and $\mathsf{Var}[X] = rp/(1-p)^2$, respectively. A negative-binomial distributed random variable $y \sim \text{NB}(r, p)$ can be generated from a gamma mixed Poisson distribution as, $y \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(r, \frac{p}{1-p})$ by marginalizing over λ.

2.2.5 *The Sum-Logarithmic distribution*

A discrete random variable X is said to have a logarithmic (Log) distribution if for $k = 1, 2, \dots,$ the probability mass function of X is defined by

$$f_X(x \mid p) = \frac{-p^x}{x \ln(1-p)}. \tag{2.2.5}$$

A random variable M is said to have a sum-logarithimic distribution $m \sim \mathrm{SumLog}(l, p)$ when $m = \sum_{t=1}^{l} u_t, u_t \sim \mathrm{Log}(p)$.

2.2.6 *The Poisson Randomized Gamma distribution*

The Poisson ramdomized gamma (PRG) distributed random variable x has the probability mass function as

$$f_X(x \mid \lambda, \beta) = [\exp(-\lambda)]^{\delta(x=0)} \left[\exp(-\lambda - \beta x)\Big(\frac{\lambda\beta}{x}\Big)^{1/2} I_{-1}\Big(2(\lambda\beta x)^{1/2}\Big)^{\delta(x>0)}\right], \tag{2.2.6}$$

where

$$I_{-1}(\alpha) = (\frac{\alpha}{2})^{-1} \sum_{n=1}^{\infty} \frac{(\frac{\alpha^2}{4})^n}{n!\Gamma(n)}, \quad \alpha > 0 \tag{2.2.7}$$

is the modified Bessel function of the first kind $I_\nu(\alpha)$ with ν fixed at -1. Using the laws of total expectation and total variance, one may show that

$$\mathsf{E}[X] = \lambda/\beta, \tag{2.2.8}$$

$$\mathsf{Var}[X] = 2\lambda/\beta^2. \tag{2.2.9}$$

A Poisson randomized gamma distributed variable $x \sim \mathrm{PRG}(\lambda, \beta)$ can be generated from a Poisson mixed gamma distribution as

$$\begin{aligned} x &\sim \mathrm{Gamma}(n, 1/\beta), \\ n &\sim \mathrm{Poisson}(\lambda). \end{aligned} \tag{2.2.10}$$

2.2.7 *Dirichlet processes and Chinese restaurant processes*

A Dirichlet process defines a distribution on random probability measures. Let Θ be a measurable space, and H a probability distribution on Θ, and α a positive scalar. A Dirichlet process is parameterized by a base measure H and a concentration parameter α. Given a finite partition $S_1, S_2, \dots, S_K$ of Θ, a random probability distribution G on Θ is a drawn from a Dirichlet process, $\mathrm{DP}(\alpha, H)$, if its measure on every finite partitions follows a Dirichlet distribution as

$$(G(S_1), G(S_2), \dots, G(S_K)) \sim \mathrm{Dirichlet}(\alpha H(S_1), \alpha H(S_2), \dots, \alpha H(S_K)).$$

Given a draw from a Dirichlet process $G \sim \mathrm{DP}(\alpha, H)$, we assume that the variables $\{\theta_i\}_{i=1}^N$ are independently sampled from G, and thus exchangeable. Marginalizing out G, the predictive distribution of θ_{N+1} conditioning on $\{\theta_i\}_{i=1}^N$ can be expressed as

$$\theta_{N+1} \mid \{\theta_i\}_{i=1}^N, \alpha, H \sim \sum_{k=1}^{K} \frac{m_k}{N+\alpha} \delta_{\psi_k} + \frac{\alpha}{N+\alpha} H, \tag{2.2.11}$$

where $\{\psi_k\}_{k=1}^K$ are distinct values taken on by $\{\theta_i\}_{i=1}^N$, and $m_k = \sum_{i=1}^N 1(\theta_i = \psi_k)$ is the number of variables that are equal to ψ_k. The stochastic process described in Eq. 2.2.11 is known as the Pólya urn scheme (Blackwell and MacQueen 1973) and also the Chinese restaurant process (Pitman 2006; Teh et al. 2007).

Under the Chinese restaurant process, the number of distinct atoms L is a random variable depending on the total number of samples n and the concentration parameter α. Let $S(n,l)$ denote unsigned Stirling numbers of the first kind, it is shown in (M. Zhou and L. Carin 2015a) that the random count variable L has the probability mass function as

$$f_L(l \mid n, \alpha) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} |S(n,l)| \alpha^l, \quad l = 0, 1, \ldots, n, \tag{2.2.12}$$

which we refer to as the Chinese restaurant distribution (CRT) in this thesis. A random variable drawn from a CRT distribution $l \sim \mathrm{CRT}(n, \alpha)$ can be generated as

$$l = \sum_{i=1}^{m} b_i, \quad b_i \sim \mathrm{Bernoulli}(\frac{\alpha}{i-1+\alpha}). \tag{2.2.13}$$

2.2.8 *The Poisson-Logarithmic bivariate distribution*

The Poisson-logarithmic bivariate distributed variable (m,l) has the probability mass function as

$$f_{M,L}(m, l \mid r, p) = \frac{S(m,l) r^l}{m!} (1-p)^r p^m,$$

where $m \in \mathbb{Z}_{>0}$ and $l = 0, 1, \ldots, m$. Note that the Poisson-logarithmic bivariate distribution can be equivalently expressed as the product of a negative-binomial and a Chinese restaurant table (CRT) distributions, and also the product of a sum-logarithmic (SumLog) and a Poisson distributions as

$$\begin{aligned}\mathrm{PoisLog}(m, l\,;\, r, p) &\equiv \mathrm{CRT}(l\,;\, m, r)\mathrm{NB}(m\,;\, r, p) \\ &\equiv \mathrm{SumLog}(m\,;\, l, p)\mathrm{Poisson}(l\,;\, -r\ln(1-p)).\end{aligned}$$

2.2.9 *Gamma processes*

The gamma process (GaP) is a *completely random measure* (CRM) (Kingman 1967) defined on the product space $\Theta \times \mathbb{R}_{>0}$ as $G \sim \mathrm{GaP}(G_0, c)$, where c is a *scale* parameter, and G_0 is a finite and continuous *base measure* over a complete separable metric space Θ, such that $G(S_k) \sim \mathrm{Gamma}(G_0(S_k), c)$ are independent gamma random variables for disjoint subsets $\{S_k\}_{k=1}^{\infty}$ of Θ. The positive Lévy measure of the gamma process can be expressed as $\nu(\mathrm{d}r) = cr^{-1}e^{-cr}\mathrm{d}r$. As a completely random measure, the gamma process can be regarded as a Poisson process on

$\Theta \times \mathbb{R}_{>0}$ with mean measure $\nu(\mathrm{d}\theta, \mathrm{d}r)$. A sample from this Poisson process consists of countably infinite atoms because $\int \int_{\Theta\times\mathbb{R}_{>0}} \nu(\mathrm{d}\theta, \mathrm{d}r) = \infty$. Thus, a sample from the gamma process can be expressed as $G = \sum_{k=1}^{\infty} r_k \delta_{\theta_k} \sim \mathrm{GaP}(G_0, c)$. More detailed information about the gamma process can be found in (Wolpert et al. 1998; Wolpert, Clyde, and Tu 2011).

2.2.10 *The thinned completely random measures framework*

Let $\Pi = \{(x_k, \theta_k, r_k)\}_{k=1}^{\infty}$ be generated by a Poisson process on the augmented product space $\mathcal{X} \times \Theta \times \mathbb{R}_{>0}$ with mean measure $\nu(\mathrm{d}x, \mathrm{d}\theta, \mathrm{d}r)$. Let $G = \sum_{k=1}^{\infty} r_k \delta_{(x_k,\theta_k)}$ be a completely random measure (CRM) on $\mathcal{X} \times \Theta \times \mathbb{R}_{>0}$, and let $\mathcal{T}$ denote the time set as the coveriate. In order to construct a family of random measures $\{G^{(t)}\}_{t\in\mathcal{T}}$ dependent on covariate values $t \in \mathcal{T}$, a set of binary random variables $b_k^{(t)}$ is generated for each point $(x_k, r_k, \theta_k) \in \Pi$ such that $p(b_k^{(t)} = 1) = P_{x_k}(t)$, where $P_x : \mathcal{T} \to [0, 1]$ denotes the thinning function which determines the probability that atom k in the global measure G appears in the local measure $G^{(t)}$ at covariate value t. Then, the set of covariate-dependent CRMs $\{G^{(t)}\}_{t\in\mathcal{T}}$ can be specified as

$$G^{(t)} = \sum_{k=1}^{\infty} b_k^{(t)} r_k \delta_{\theta_k}, \qquad t \in \mathcal{T}.$$

The new CRMs are well-defined by the mapping theorem for the Poisson processes (Kingman 1993), that is proved in (Foti et al. 2013). As a concrete example, a thinned gamma process (tGaP) can be constructed to model the global atoms and their *activity/inactivity* at multiple time points originally developed for dynamic topic models. Let $\nu(\mathrm{d}x, \mathrm{d}\theta, \mathrm{d}r) = H(\mathrm{d}x)G_0(\mathrm{d}\theta)\nu_0(\mathrm{d}r)$, where $\nu_0(\mathrm{d}r) = cr^{-1}e^{-cr}\mathrm{d}r$ is the Lévy measure of the gamma process. We transform a Gaussian basis kernel pointwise using a logistic function as the thinning function:

$$P_{x_k}(t) = \sigma\left\{ \omega_{0k} + \sum_{l=1}^{T} \omega_{lk} \exp[-\phi_k (t-l)^2] \right\},$$

where $\sigma(x) = 1/(1 + \exp(-x))$ denotes the logistic function. We fix the centres of these kernels to the T discrete time points in covariate space $\mathcal{T}$. We characterize each location $x_k \in \mathcal{X}$ by a set of $T + 1$ kernel weights $\omega_{lk} \in \mathbb{R}$, and a (shared) kernel width ϕ_k uniformly drawn from a fixed dictionary $\{\phi_1^*, \ldots, \phi_D^*\}$ of size D. To encourage sparsity of the kernel weights, we place a normal-inverse gamma prior over ω_{lk}, i.e., $\omega_{lk} \sim \mathcal{NIG}(\omega_{lk}; 0, 1, 1)$. Hence, the base measure $H(dx)$ can be expressed as $H(dx) = \mathcal{NIG}(\omega_{lk}; 0, 1, 1)\mathrm{Categorical}(\phi_k; \{\phi_1^*, \ldots, \phi_D^*\})$. The generative procedure can be expressed as

$$\begin{aligned} G &= \sum_{k=1}^{\infty} r_k \delta_{(x_k,\theta_k)} \sim \mathrm{CRM}(\nu(\mathrm{d}x, \mathrm{d}\theta, \mathrm{d}r)), \\ \omega_{lk} &\sim \mathcal{NIG}(0, 1, 1), \quad \phi_k \sim \mathrm{Categorical}(\phi_1^*, \ldots, \phi_D^*), \\ P_{x_k}(t) &= \sigma\left\{ \omega_{0k} + \sum_{l=1}^{T} \omega_{lk} \exp[-\phi_k (t-l)^2] \right\}, \\ b_k^{(t)} &\sim \mathrm{Bernoulli}\Big[P_{x_k}(t)\Big], \\ G^{(t)} &= \sum_{k=1}^{\infty} b_k^{(t)} r_k \delta_{\theta_k}. \end{aligned} \tag{2.2.14}$$

2.2.11 *The Negative-Binomial augmentation scheme*

Let $r. = \sum_i r_i$, and $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ be a Dirichlet-multinomial distributed random vector: $\{x_1, \ldots, x_N\} \sim \mathrm{DirMult}(x., r_1, \ldots, r_N)$ where $x. \sim \mathrm{NB}(r., p)$, we can equivalently sample $\{x_i\}_{i=1}^N$ as $x_i \sim \mathrm{NB}(r_i, p)$, for $i = 1, \ldots, N$. Hence, given Dirichlet-Multinomial distributed random variables $\{x_1, \ldots, x_N\} \sim \mathrm{DirMult}(x., r_1, \ldots, r_N)$ where $r_i \sim \mathrm{Gamma}(a_i, b_i)$, our aim is to sample the parameter $\{r_i\}_{i=1}^N$. To this end, we can introduce an auxiliary variable p as $p \sim \mathrm{Beta}(x., r.)$, and then we have $x_i \sim \mathrm{NB}(r., p)$. We further augment x_i with a CRT distributed random variable $l_i \sim \mathrm{CRT}(x_i, r_i)$, and then according to the Poisson-Logarithmic bivariate distribution, we obtain

$$
\begin{aligned}
x_i &\sim \mathrm{SumLog}(p), \\
l_i &\sim \mathrm{Poisson}[-r_i \log(1-p)].
\end{aligned}
$$

Then, the conditional distribution of r_i can be easily obtained using the gamma-Poisson conjugacy.

2.2.12 *Hawkes processes*

Let $N(t)$ be a counting process recording the number of events occurring at times $\{t_i\}$ with $t_i < t$. The probability of an event occurring in a small time interval $[t, t + \mathrm{d}t)$ is given by $\Pr(\mathrm{d}N(t) = 1 \mid \mathcal{H}(t)) = \lambda(t)\mathrm{d}t$, where $\mathcal{H}(t) \equiv \{t_i \mid t_i < t\}$ denotes the history of events up to but not including time t, $\mathrm{d}N(t)$ is the increment of the process, and $\lambda(t)$ is the conditional intensity function (intensity, for short) of $N(t)$. A Hawkes process is a doubly-stochastic point process (Daley et al. 2003) with the intensity function

$$
\begin{aligned}
\lambda(t) &= \mu + \int_0^t \gamma(t-s)\mathrm{d}N(s) \\
&= \mu + \sum_{j:t_j \in \mathcal{H}(t)} \gamma(t - t_j),
\end{aligned} \tag{2.2.15}
$$

where $\mu \geq 0$ is the base rate capturing the *exogenous* activities, and $\gamma(t)$ is the nonnegative triggering kernel modelling the *endogenous* activities. Note that this intensity function characterizes the self-excitation effects that past events have on the current event rate. Here, we consider an exponential kernel $\gamma(t-s) \equiv \alpha \exp[-(t-s)/\delta]$ where $\alpha \geq 0$ determines the magnitude of excitations, which exponentially decays with a constant rate $\delta \geq 0$. The stationary condition for Hawkes processes requires $\alpha\delta < 1$. Recent work (Blundell, Beck, and Heller 2012; J. Yang et al. 2017; Miscouridou et al. 2018) has been proposed to capture the reciprocity in communications between a pair of individuals using mutually-exciting Hawkes processes. Formally, for a pair of nodes $u, v \in V$, we have the counting process $N_{uv}(t)$, which defines the number of *directed* interactions from node u to node v in the time interval $[0, t)$. Let the history of interactions from nodes u to v be denoted as $\mathcal{H}_{uv}(t)$. Accordingly, $N_{uv}(t)$ and $N_{vu}(t)$ are mutually-exciting Hawkes processes if their intensity functions take the forms

$$
\begin{aligned}
\lambda_{uv} &= \mu_{uv} + \sum_{t_j \in \mathcal{H}_{vu}(t)} \gamma(t - t_j), \\
\lambda_{vu} &= \mu_{vu} + \sum_{t_i \in \mathcal{H}_{uv}(t)} \gamma(t - t_i),
\end{aligned} \tag{2.2.16}
$$

respectively. Note that mutually-exciting Hawkes processes capture the reciprocating interactions from node u to node v at time t as a response to the past interactions from v to u, and vice versa.

2.3 NETWORK MODELS

2.3.1 *Static network models*

A large amount of contributions have been dedicated to analyzing static networks. These static network analysis methods include the modularity maximization method (Newman and Girvan 2004), clique percolation (Palla et al. 2005), matrix and tensor factorization based methods (Dunlavy et al. 2011), and various probabilistic methods (Airoldi et al. 2008; Miller et al. 2009; P. D. Hoff et al. 2001). See (Goldenberg et al. 2010) for a comprehensive review. In this thesis, we focus on the probabilistic methods as these methods not only can capture the global network properties via appropriate prior specifications, but also can be applied to detect underlying community structure and to predict missing edges. The statistical models for static networks can largely be classified as class-based and feature-based models. Class-based statistical models for static networks, such as stochastic block models (Holland et al. 1983; Nowicki and Snijders 2001), assume that a static network can be decomposed into a finite number of latent communities that nodes can belong to, and that these communities entirely determine the formation of network edges. Therefore, inference in the class-based models reduces to assigning each node $v \in \mathcal{V}$ to one of the latent communities $k = 1, \ldots, K$, and estimating the community interaction matrix B of size $K \times K$. For two nodes u and v, their latent community assignments are denoted as z_u and z_v, respectively. Thus, the edge probability between u and v take the form

$$A_{uv} \sim \text{Bernoulli}(B_{z_u z_v}). \tag{2.3.1}$$

The stochastic block model (Holland et al. 1983; Nowicki and Snijders 2001) requires the number of latent communities to be determined in advance. The infinite relational model (IRM) (Kemp et al. 2006) extends the stochastic block model by generating the node-community memberships from a Chinese restaurant process prior, and thus allows the number of communities to be inferred in a Bayesian nonparametric way. Moreover, stochastic block models fail to capture the overlapping community structure of networks. To tackle this issue, the mixed membership stochastic block model (MMSBM) (Airoldi et al. 2008) is developed to allow each node to have mixed memberships. More specifically, we endow each node $u \in \mathcal{V}$ with a latent mixed memberships vector π_u, which is drawn from the prior $\pi_u \sim \text{Dirichlet}(\alpha)$, where α denotes the hyperparameter. For nodes u and v, node u draws one role $z_{u\to v} \sim \text{Multinomial}(\pi_u)$ as a sender of the interaction with node v, and node v draws one role $z_{u\leftarrow v} \sim \text{Multinomial}(\pi_v)$ as a receiver of the interaction from node u. Finally, the edge probability from u to v is

$$\begin{aligned} z_{u\to v} &\sim \text{Multinomial}(\pi_u), \\ z_{u\leftarrow v} &\sim \text{Multinomial}(\pi_v), \\ A_{uv} &\sim \text{Bernoulli}(z_{u\to v} B z_{u\leftarrow v}). \end{aligned} \tag{2.3.2}$$

The mixed membership stochastic block model captures the node-community memberships using a multinomial distributed representation. Thus, the MMSBM assumes that by increasing the degree of a node's membership to some group k, the same node's degrees to the other groups $k' \neq k$ have to decrease as the same node's memberships normalize to one. To circumvent this limitation, the latent feature relational model (LFRM) is developed to represent each node $u \in \mathcal{V}$ by a set of binary features that govern the edge formation of node u with the other nodes of the network. In particular, the LFRM utilizes the Indian buffet process as the prior for the node-memberships, and captures the feature interactions using a real-valued weight matrix $W \in \mathbb{R}^{K \times K}$. The probability of an edge between node u and node v take the form

$$\Pr(A_{uv} = 1 \mid -) = \sigma(\sum_{k,k'} z_{uk} w_{kk'} z_{vk'}), \tag{2.3.3}$$

where $\sigma(x) \equiv \frac{1}{1+\exp(-x)}$ is the logistic link function. Despite showing expressiveness, the latent feature relational model fails to capture the degrees of each node's memberships to different communities only using binary feature vectors.

2.3.2 *Models for discrete-time networks*

Real-world network data, such as friend relationships or interactions on online social networks, is often dynamic because these observed relations among nodes may appear or disappear over time. A dynamic network data can be composed by taking a series of snapshots at multiple time points. Although a large amount of effort has been dedicated to studying dynamic networks, we review some representative methods for discrete-time dynamic networks. Since the stochastic block model is widely used in analysis of static networks, many previous work (Fu et al. 2009; Xing et al. 2010; Ho et al. 2011; K. S. Xu et al. 2014) extend the stochastic block model for dynamic context. These dynamic stochastic block models generally assume that the latent community memberships of each node evolve over time, and thus the edges of the same node to the other nodes change accordingly. To capture such temporal dynamics, the dynamic stochastic block model utilizes the hidden Markov model to evolve each node membership over time, and then generates each network snapshot with a stochastic block model framework.

The main difference between the previous statistical dynamic network models lies in how to capture the evolving behaviours of nodes over time. Among these methods, the dynamic latent feature relational models (J. R. Foulds et al. 2011; Heaukulani et al. 2013; M. Kim et al. 2013) employ the factorial hidden Markov model (Zoubin Ghahramani et al. 1996) to characterize the temporal dynamics of the latent node-community memberships. Other representative models include the dynamic Euclidean latent space model (Sarkar et al. 2006) that describes each node with a latent position in Euclidean latent space. To capture temporal dynamics of networks, the dynamic Euclidean latent space model allows the latent position of each node to smoothly move over time.

2.3.3 *Models for continuous-time networks*

In many cases, temporal relational events among entities are continuously generated, and thus each of these events is associated with a timestamp. Modelling timestamped relational events enable us to not only discover the implicit social structure, but also to capture the causal relationships between these events.

Some prior models (Blundell, Beck, and Heller 2012; Du et al. 2015b; H. Xu et al. 2016a; J. Yang et al. 2017; Miscouridou et al. 2018) have been developed to combine a static network model that characterizes the underlying community structure behind interactions, and the temporal point process (Hawkes 1971) that captures reciprocating interaction dynamics. For instance, the Hawkes process infinite relational model (Hawkes-IRM) (Blundell, Beck, and Heller 2012) extends the infinite relational model into a continuous-time model by characterizing interactions between each pair of two nodes using a mutually-exciting Hawkes process with the base intensity determined by the respective unique communities of these two nodes. Moreover, the Hawkes compound completely random measure model (Hawkes-CCRM) (Miscouridou et al. 2018) hybridies the compound completely random measure model that captures the global sparsity and degree heterogeneity of networks, with mutually-exciting Hawkes processes for modelling reciprocity in interactions. Other models include Hawkes dual latent space model (J. Yang et al. 2017) that embeds the nodes of a network into Euclidean space, and that builds the interaction intensity between two nodes on their latent positions. In addition, the work of (Tan et al. 2018) employs the nested Chinese restaurant process (NCRP) (D. M. Blei et al. 2010) to induce an hierarchical structure embedded with Hawkes processes for temporal interaction events.

2.3.4 *Models for networks and nodes features*

Many prior work (Rai 2017; D. I. Kim et al. 2012; P. K. Gopalan et al. 2014; Hu et al. 2016a) have considered to incorporate node features into network models to predict missing edges. Some of the work (D. I. Kim et al. 2012; P. K. Gopalan et al. 2014) leverage the given relational graph and partially observed node labels to predict the labels for the other unlabeled nodes. These methods for joint modelling of a network and its associated node features include regression models and generative models. The regression models (Rai 2017) generally treat node features as input covariates and the observed network edges as predictors. Then, predicting missing edges can be performed under the regression model using the available node features.

The generative models (D. I. Kim et al. 2012; P. K. Gopalan et al. 2014) typically capture the joint distribution of node features and network edges by factorizing the adjacency matrix and covariates in the common latent space.

3

DISCRETE TIME DYNAMIC NETWORK MODELS

The study of relational data arising from various networks including social, biological and physical networks is becoming increasingly important due to the emergence of massive relational data collected from these domains. Many efforts have been dedicated to developing statistical models in terms of community detection and missing link prediction for static networks, where either a single snapshot of the network of interest or an aggregated network over time is presented. However, network data, such as friendships or interactions in a social network, is often dynamic since the relations among the entities within the network may appear or disappear over time (Mucha et al. 2010). Hence, appropriate models are needed to enable a better understanding of the formation and evolution of dynamic networks (Phan and Airoldi 2015).

In this chapter, a probabilistic framework is developed to model such dynamic networks assuming the network of interest is composed of a set of latent communities. Each node of the observed network is hence associated with a time-dependent memberships vector that governs its involvement in multiple communities and interactions with other nodes. The node-community memberships are assumed to be gamma distributed, thus, naturally *nonnegative* real-valued. Moreover, to capture time-evolving interactions between groups of nodes, we also extend the developed dynamic network model to capture the *birth* and *death* dynamics of individual communities explicitly via a dependent relational gamma process (dRGaP). The ideal number of latent communities can be learned from data via the shrinkage mechanism of the dRGaP.

Explicitly modelling community birth/death dynamics can be useful in many applications. For instance, latent communities in a network of military disputes between countries could mean alliances such as the North Atlantic Treaty Organization (NATO) coordinating collective defence to attacks by external forces. These communities can be born and die afterwards. For example, the Warsaw Pact was established during the Cold War and dissolved in later years. We demonstrate that the proposed model can discover interpretable latent structure on a real network of military interstate disputes (Ghosn et al. 2004) that agrees with our knowledge of international relations. Furthermore, it is reasonable to model the time-evolving memberships of each individual node to interpret its joining and withdrawing behavior to these communities. Hence, we capture the dynamics of individual node-community memberships evolving over time via gamma Markov processes. In contrast to dynamic network modelling using logistic or probit mapping functions (J. R. Foulds et al. 2011; Heaukulani et al. 2013; Durante et al. 2014), we leverage the Bernoulli Poisson link (BPL) function (Dunson et al. 2005; M. Zhou et al. 2015) to generate edges from the latent space representation, which makes the computational cost of the proposed model to scale linearly with the number of edges, rather than quadratically with the number of nodes. In addition, the BPL function is also a more appropriate model for *imbalanced* binary data (Hu et al. 2015), which makes the proposed model appealing for analyzing real-world relational data that are usually extremely sparse. To perform inference, we present an efficient Gibbs sampling algorithm exploiting the Pólya-gamma data augmentation technique (Polson et al. 2013) and the data augmentation and marginalization technique for discrete data (M. Zhou and L. Carin 2015b).

The rest of this chapter is organized as follows. In Section 3.1, some related work are discussed. Section 3.2 presents the dynamic Poisson gamma membership model. In Section 3.3, we extend the dynamic Poisson gamma membership model to capture the birth and death dynamics of latent communities using the thinned completely random measure framework. Section 3.4 presents the experimental results of the two developed models compared with state-of-the-art methods on both synthetic and real-world datasets. Section 3.5 summarizes this chapter.

3.1 RELATED WORK

Prior works on discrete-time dynamic networks modelling include the exponential random graph model (ERGM) (Guo et al. 2007), matrix and tensor factorization based methods (Dunlavy et al. 2011) and statistical models (Sarkar et al. 2006; Ishiguro et al. 2010; Durante and Dunson 2014). Statistical dynamic network models received considerable attention because these models have favourable interpretability by providing uncertainty estimates for the uncovered latent representations (P. D. Hoff et al. 2001). Dynamic extensions of the mixed membership stochastic blockmodel (MMSB) (Airoldi et al. 2008) have been developed (Fu et al. 2009; Xing et al. 2010; Ho et al. 2011) using linear state space models to capture the evolution of *real-valued* node-community memberships. We note that a mixed memberships model can be recovered from our proposed model by the normalization of the gamma distributed memberships to restrict probabilities normalized to one because of the relationship between the Dirichlet and gamma distribution (Ferguson 1973). Recently, an extended Kalman filter (EKF) based algorithm (K. S. Xu et al. 2014) was proposed to infer dynamic stochastic blockmodels (SBM) with competitive performance. Dynamic extensions of the latent feature relational model (LFRM) (Miller et al. 2009) using an infinite factorial hidden Markov process to capture the evolution of *binary* node-community memberships include the dynamic relational infinite feature model (DRIFT) (J. R. Foulds et al. 2011), the latent feature propagation model (LFP) (Heaukulani et al. 2013), and the dynamic multi-group membership graph model (DMMG) (M. Kim et al. 2013).

The proposed framework is a form of bilinear Poisson factorization model (M. Zhou and L. Carin 2015b; P. Gopalan et al. 2015), and can be considered as the dynamic extension of the hierarchical gamma process edge partition model (HGP-EPM) (M. Zhou et al. 2015). The dependent completely random measure (CRM) framework (Foti et al. 2013) has been exploited for dynamic topic models and dependent latent feature models previously. To the best of our knowledge, this is the first attempt to model activity and inactivity of latent communities using a thinned CRMs framework in dynamic networks modelling. Our Markov chain construction, used to capture the evolution of node-community memberships, is inspired by the data augmentation technique (M. Zhou and L. Carin 2015b) that has been exploited for dynamic matrix factorization (Acharya et al. 2015a; Schein et al. 2016b) and deep gamma belief networks (M. Zhou et al. 2016). We note that the dynamic gamma process Poisson factorization (D-GPPF) (Acharya et al. 2015b) has been proposed using gamma Markov chains to model the evolution of latent communities while the D-GPPF assumes node memberships are static over time. Notably, a gamma Markov chain can alternatively be constructed conditioning the state at every time step on the state at previous time step via gamma scale parameters (Ranganath et al. 2015).

3.2 THE DYNAMIC POISSON GAMMA MEMBERSHIPS MODEL

In the proposed model, each node $u \in \mathcal{V}$ is characterized by a time-dependent latent membership variable $\phi_{uk}^{(t)}$ that determines its interactions or involvement in community k at the t-th snapshot of the dynamic networks. This latent node-community membership is modeled by a gamma random variable and is, thus, naturally *nonnegative real-valued.* This is contrast to *multi-group memberships* models (or latent feature relational models) (J. R. Foulds et al. 2011; Heaukulani et al. 2013; M. Kim et al. 2013) where each node-community membership is represented by a *binary* latent feature indicator. The multi-group memberships models assume that each node either associates to one community or not – simply by a binary feature indicator. The proposed model on the other hand can characterize how strongly each node associates with multiple communities.

3.2.1 *Dynamics of latent node-community memberships.*

For dynamic networks, the latent node-community membership $\phi_{uk}^{(t)}$ can evolve over time to interpret the interaction dynamics among the nodes. For example, latent community k could mean "play soccer" and $\phi_{uk}^{(t)}$ could mean how frequently person u plays soccer or how strongly person u likes playing soccer. The person's degree of association to this community could be increasing over time due to, for instance, increased interaction with professional soccer players, or decreasing over time as a consequence of sickness. Hence, in order to model the temporal evolution of the latent node-community memberships, we assume the individual memberships to form a gamma Markov chain. More specifically, $\phi_{uk}^{(t)}$ is drawn from a gamma distribution, whose shape parameter is the latent membership at the previous time

$$\begin{aligned} \phi_{uk}^{(t)} &\sim \text{Gamma}(\phi_{uk}^{(t-1)}/\tau, 1/\tau), \qquad \text{for } t = 1, \dots, T \\ \phi_{uk}^{(0)} &\sim \text{Gamma}(g_0, 1/h_0), \end{aligned}$$

where the parameter τ controls the variance without affecting the mean, i.e.,
$\mathsf{E}[\phi_{uk}^{(t)} \mid \phi_{uk}^{(t-1)}, \tau] = \phi_{uk}^{(t-1)}$.

3.2.2 *Model of latent communities.*

The interactions or correlations among latent communities are characterized by a matrix Ω of size $K \times K$, where $\Omega_{kk'}$ relates to the probability of there being a link between node u affiliated to community k and node v affiliated to community k'. Specifically, we assume the latent communities to be generated by the following hierarchical process: we first generate a separate weight for each community as

$$r_k \sim \text{Gamma}(\gamma_0/K, 1/c_0), \tag{3.2.1}$$

and then generate the inter-community interaction weight $\Omega_{kk'}$ and intra-community weight Ω_{kk} as

$$\Omega_{kk'} \sim \begin{cases} \text{Gamma}(\xi r_k, 1/\beta), & \text{if } k = k' \\ \text{Gamma}(r_k r_{k'}, 1/\beta), & \text{otherwise} \end{cases} \tag{3.2.2}$$

where $\xi \in \mathbb{R}_{>0}$ and $\beta \in \mathbb{R}_{>0}$. The reasonable number of latent communities can be inferred from dynamic relational data itself by the *shrinkage* mechanism of the proposed model. More specifically, for fixed γ_0, the redundant communities will effectively be shrunk as many of the communities weights tend to be small for increasing K. Thus, the interaction weights $\Omega_{kk'}$ between the redundant community k and all the other communities k', and all the node-memberships to community k will be shrunk accordingly. In practice, the intra-community weight Ω_{kk} would tend to almost zero if $\Omega_{kk} \sim \mathrm{Gamma}(r_k^2, 1/\beta)$ for small r_k, and the corresponding communities will disappear inevitably. Hence, we use a separate variable ξ to avoid overly shrinking of small communities with less interactions with other communities. As γ_0 has a large effect on the number of the latent communities, we do not treat it as a fixed parameter but place a gamma prior over it, i.e., $\gamma_0 \sim \mathrm{Gamma}(1, 1)$. Given the latent node-community membership $\phi_{uk}^{(t)}$ and the interaction weights $\Omega_{kk'}$ among communities, the probability of there being a link between node u and v is given by

$$A_{uv}^{(t)} \sim \mathrm{Bernoulli}\left(1 - \exp\left\{-\sum_{k=1}^{K}\sum_{k'=1}^{K} \Omega_{kk'}\phi_{uk}^{(t)}\phi_{vk'}^{(t)}\right\}\right). \tag{3.2.3}$$

Interestingly, we can also generate $A_{uv}^{(t)}$ by truncating a latent count random variable $\tilde{A}_{uv}^{(t)}$ at 1, where $\tilde{A}_{uv}^{(t)}$ can be seen as the integer-valued *weight* for node u and v, and can be interpreted as the number of times the two nodes interacted. More specifically, $A_{uv}^{(t)}$ can be drawn as

$$A_{uv}^{(t)} = \mathbb{1}(\tilde{A}_{uv}^{(t)} \geq 1), \tag{3.2.4}$$

$$\tilde{A}_{uv}^{(t)} \sim \mathrm{Poisson}(\sum_{k=1}^{K}\sum_{k'=1}^{K} \Omega_{kk'}\phi_{uk}^{(t)}\phi_{vk'}^{(t)}). \tag{3.2.5}$$

We can obtain (3.2.3) by marginalizing out the latent count $\tilde{A}_{uv}^{(t)}$ from the above expression. The conditional distribution of the latent count $\tilde{A}_{uv}^{(t)}$ can then be written as

$$(\tilde{A}_{uv}^{(t)} \mid A_{uv}^{(t)}, \Phi, \Omega) \sim A_{uv}^{(t)}\mathrm{Poisson}_{+}(\sum_{k=1}^{K}\sum_{k'=1}^{K} \Omega_{kk'}\phi_{uk}^{(t)}\phi_{vk'}^{(t)}),$$

where $x \sim \mathrm{Poisson}_{+}(\sigma)$ is the zero-truncated Poisson distribution with probability mass function (PMF) $f_X(x|\sigma) = (1 - e^{-\sigma})^{-1}\sigma^x e^{-\sigma}/x!$, $x \in \mathbb{Z}_{>0}$, and Φ denotes the set of all node-community membership variables. The usefulness of this construction for $A_{uv}^{(t)}$ will become clear in the inference section. We note that the latent count $\tilde{A}_{uv}^{(t)}$ only needs to be sampled for $A_{uv}^{(t)} = 1$, using rejection sampling detailed in (M. Zhou et al. 2015). The proposed hierarchical generative model is as follows:

$$\begin{aligned}
\phi_{uk}^{(t)} &\sim \mathrm{Gamma}(\phi_{uk}^{(t-1)}/\tau, 1/\tau), \ \text{ for } t = 1, \ldots, T \\
\phi_{uk}^{(0)} &\sim \mathrm{Gamma}(g_0, 1/h_0), \\
r_k &\sim \mathrm{Gamma}(\gamma_0/K, 1/c_0), \\
\Omega_{kk'} &\sim \begin{cases} \mathrm{Gamma}(\xi r_k, 1/\beta), & \text{if } k = k' \\ \mathrm{Gamma}(r_k r_{k'}, 1/\beta), & \text{otherwise} \end{cases} \\
\tilde{A}_{uv}^{(t)} &\sim \mathrm{Poisson}(\sum_{k=1}^{K}\sum_{k'=1}^{K} \phi_{uk}^{(t)}\Omega_{kk'}\phi_{vk'}^{(t)}), \\
A_{uv}^{(t)} &= \mathbb{1}(\tilde{A}_{uv}^{(t)} \geq 1).
\end{aligned}$$

For the model's hyperparameters, we draw c_0, ξ and β from $\mathrm{Gamma}(0.1, 0.1)$. The graphical model is shown in Fig. 3.1.

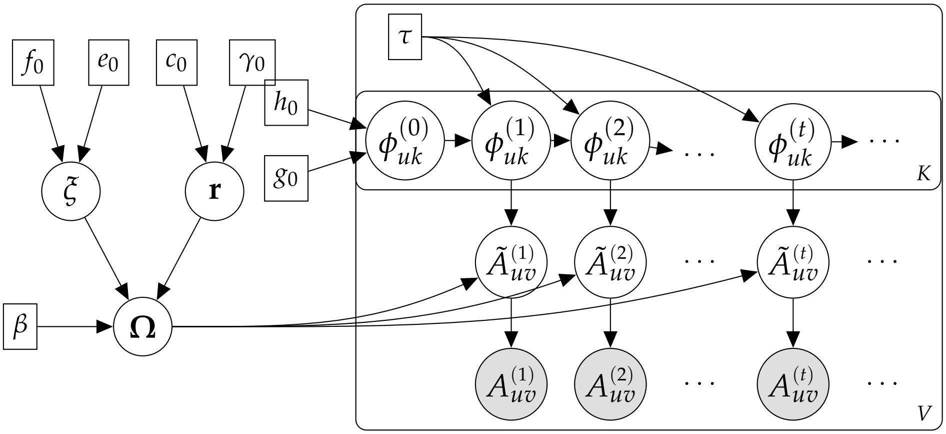


Figure 3.1: The graphical model of the dynamic Poisson gamma memberships model; auxillary variables introduced for inference are not shown.

3.2.3 *Inference algorithm*

A Gibbs sampling algorithm for the dynamic Poisson gamma memberships model is derived to draw samples of the model parameters $\{\phi^{(t)}_{uk}, \Omega_{kk'}, r_k, \xi, \gamma_0, \beta, c_0\}$ from their conditional posterior distribution given the observed dynamic relational data and the hyper-parameters $\{e_0, f_0, g_0, h_0\}$. In order to circumvent the technical challenging of drawing samples from the gamma Markov chain which does not have a closed-form conditional posterior, we generalize the statistical ideas of data augmentation and marginalization technique and the gamma-Poisson conjugacy to derive a closed-form update.

Sampling latent count $\tilde{A}^{(t)}_{uv}$. We sample a latent count for each time dependent observed edge $A^{(t)}_{uv}$ as

$$(\tilde{A}^{(t)}_{uv}|-) \sim A^{(t)}_{uv}\text{Poisson}_{+}(\sum_{k=1}^{K}\sum_{k'=1}^{K}\Omega_{kk'}\phi^{(t)}_{uk}\phi^{(t)}_{vk'}). \tag{3.2.6}$$

Sampling $\tilde{A}^{(t)}_{ukk'v}$. Using the Poisson additive property, we can augment the latent count $\tilde{A}^{(t)}_{uv} \sim \text{Poisson}(\sum_{k=1}^{K}\sum_{k'=1}^{K}\Omega_{kk'}\phi^{(t)}_{uk}\phi^{(t)}_{vk'})$ as $\tilde{A}^{(t)}_{uv} = \sum_{k,k'=1}^{K}\tilde{A}^{(t)}_{ukk'v}$, where $\tilde{A}^{(t)}_{ukk'v} \sim \text{Poisson}(\Omega_{kk'}\phi^{(t)}_{uk}\phi^{(t)}_{vk'})$. Then, via the Poisson-multinomial equivalence, we partition the latent count $\tilde{A}^{(t)}_{uv}$ as

$$(\tilde{A}^{(t)}_{ukk'v}|-) \sim \text{Multinomial}(\tilde{A}^{(t)}_{uv}; \frac{\Omega_{kk'}\phi^{(t)}_{uk}\phi^{(t)}_{vk'}}{\sum_{k=1}^{K}\sum_{k'=1}^{K}\Omega_{kk'}\phi^{(t)}_{uk}\phi^{(t)}_{vk'}}). \tag{3.2.7}$$

Sampling r_k. Via the Poisson additive property, we have

$$(\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}|-) \sim \text{Poisson}(\Omega_{kk'}\theta_{kk'}), \tag{3.2.8}$$

where $\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} = \sum_t\sum_{u,v\neq u}\tilde{A}^{(t)}_{ukk'v}$ and $\theta_{kk'} = \sum_t\sum_v\sum_{u\neq v}\phi^{(t)}_{uk}\phi^{(t)}_{vk'}$.

We can marginalize out $\Omega_{kk'}$ from (3.2.8) and (3.2.2) using the gamma-Poisson conjugacy, and then have

$$(\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}|-) \sim \mathrm{NB}(r_k \xi^{\delta(kk')} (r_{k'})^{1-\delta(kk')}, \tilde{p}_{kk'}),$$

where $\tilde{p}_{kk'} = \frac{\theta_{kk'}}{\theta_{kk'}+\beta}$ and the delta function $\delta(kk') = 1$ if $k = k'$.

Exploiting the Poisson logarithmic bivariate distribution, we introduce the CRT distributed auxiliary variables as

$$l_{kk'} \sim \mathrm{CRT}(\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}, r_k \xi^{\delta(kk')} (r_{k'})^{1-\delta(kk')}). \tag{3.2.9}$$

We then re-express the bivariate distribution over $\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}$ and $l_{kk'}$ as

$$\begin{aligned}(\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}|-) &\sim \mathrm{SumLog}(l_{kk'}, r_k \xi^{\delta(kk')} (r_{k'})^{1-\delta(kk')}),\\ (l_{kk'} \mid -) &\sim \mathrm{Poisson}(-r_k \xi^{\delta(kk')} (r_{k'})^{1-\delta(kk')} \ln(1-\tilde{p}_{kk'})).\end{aligned} \tag{3.2.10}$$

Using (3.2.1) and (3.2.10), via the gamma-Poisson conjugacy, we obtain the conditional distribution of r_k as

$$(r_k \mid -) \sim \mathrm{Gamma}\Big[\frac{\gamma_0}{K} + \sum_{k'} l_{kk'}, \frac{1}{c_0 - \sum_{k'} \xi^{\delta(kk')} (r_{k'})^{1-\delta(kk')} \ln(1-\tilde{p}_{kk'})}\Big]. \tag{3.2.11}$$

Sampling ξ. We resample the auxiliary variables l_{kk} using (3.2.9), and then exploit the gamma-Poisson conjugacy to sample ξ as

$$(\xi \mid -) \sim \mathrm{Gamma}\Big[e_0 + \sum_k l_{kk}, \frac{1}{f_0 - \sum_k r_k \ln(1-\tilde{p}_{kk})}\Big]. \tag{3.2.12}$$

Sampling $\Omega_{kk'}$. We sample $\Omega_{kk'}$ from its conditional distribution obtained using (3.2.2) and (3.2.8) via the gamma-Poisson conjugacy as

$$(\Omega_{kk'}|-) \sim \mathrm{Gamma}\Big[\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} + r_k \xi^{\delta(kk')} (r_{k'})^{1-\delta(kk')}, 1/(\beta + \theta_{kk'})\Big]. \tag{3.2.13}$$

Sampling γ_0. Using (3.2.10) and the Poisson additive property, we have $l_{k\cdot} = \sum_{k'} l_{kk'}$ as

$$(l_{k\cdot}|-) \sim \mathrm{Poisson}(-r_k \sum_{k'} \xi^{\delta(k,k')} (r_{k'})^{1-\delta(k,k')} \ln(1-\tilde{p}_{kk'})).$$

Marginalizing out r_k using the gamma-Poisson conjugacy, we have

$$(l_{k\cdot}|-) \sim \mathrm{NB}(\gamma_0/K, \hat{p}_k),$$

where $\hat{p}_k = \frac{\sum_{k'} \xi^{\delta(k,k')} (r_{k'})^{1-\delta(k,k')} \ln(1-\tilde{p}_{kk'})}{c_0 - \sum_{k'} \xi^{\delta(k,k')} (r_{k'})^{1-\delta(k,k')} \ln(1-\tilde{p}_{kk'})}$. We introduce the auxiliary variables $\tilde{l}_k \sim \mathrm{CRT}(l_{k\cdot}, \gamma_0/K)$, and re-express the bivariate distribution over $l_{k\cdot}$ and $\tilde{l}_k$ as

$$\begin{aligned}(l_{k\cdot} \mid -) &\sim \mathrm{SumLog}(\tilde{l}_k, \hat{p}_k),\\ (\tilde{l}_k \mid -) &\sim \mathrm{Poisson}(-\frac{\gamma_0}{K} \ln(1-\hat{p}_k)).\end{aligned} \tag{3.2.14}$$

Then, via the gamma-Poisson conjugacy, we sample γ_0 as

$$(\gamma_0 \mid -) \sim \text{Gamma}\Big[1 + \sum_k \tilde{l}_k, \frac{1}{1 - \frac{1}{K}\sum_k \ln(1 - \hat{p}_k)}\Big]. \tag{3.2.15}$$

Sampling latent memberships $\phi_{uk}^{(t)}$. Since the latent memberships $\phi_{uk}^{(t)}$ evolve over time according to our Markovian construction, the backward and forward information need to be incorporated into the updates of $\phi_{uk}^{(t)}$. We start from time slice $t = T$,

$$\begin{aligned} \tilde{A}_{uk\cdot\cdot}^{(T)} &\sim \text{Poisson}(\phi_{uk}^{(T)} \omega_{uk}^{(T)}), \\ \phi_{uk}^{(T)} &\sim \text{Gamma}(\phi_{uk}^{(T-1)}/\tau, 1/\tau), \end{aligned}$$

where

$$\begin{aligned} \tilde{A}_{uk\cdot\cdot}^{(t)} &\equiv \sum_{v \neq u, k'} \tilde{A}_{ukk'v}^{(t)}, \\ \omega_{uk}^{(t)} &\equiv \sum_{v \neq u, k'} \phi_{vk'}^{(t)} \Omega_{kk'}. \end{aligned}$$

Via the gamma-Poisson conjugacy, we have

$$(\phi_{uk}^{(T)} \mid -) \sim \text{Gamma}\Big[\phi_{uk}^{(T-1)}/\tau + \tilde{A}_{uk\cdot\cdot}^{(T)}, 1/(\tau + \omega_{uk}^{(T)})\Big]. \tag{3.2.16}$$

Marginalizing out $\phi_{uk}^{(T)}$ yields

$$\tilde{A}_{uk\cdot\cdot}^{(T)} \sim \text{NB}(\phi_{uk}^{(T-1)}/\tau, \varrho_{uk}^{(T)}), \tag{3.2.17}$$

where $\varrho_{uk}^{(T)} = \frac{\omega_{uk}^{(T)}}{\tau + \omega_{uk}^{(T)}}$.

Exploiting the Poisson logarithmic bivariate distribution, the NB distribution can be augmented with a CRT distributed auxiliary variable as

$$\begin{aligned} \tilde{A}_{uk\cdot\cdot}^{(T)} &\sim \text{NB}(\phi_{uk}^{(T-1)}/\tau, \varrho_{uk}^{(T)}), \\ \hat{m}_{uk}^{(T)} &\sim \text{CRT}(\tilde{A}_{uk\cdot\cdot}^{(T)}, \phi_{uk}^{(T-1)}/\tau). \end{aligned}$$

Then, we re-express the bivariate distribution over $\tilde{A}_{uk\cdot\cdot}^{(T)}$ and $\hat{m}_{uk}^{(T)}$ as

$$\begin{aligned} \tilde{A}_{uk\cdot\cdot}^{(T)} &\sim \text{SumLog}(\hat{m}_{uk}^{(T)}, \varrho_{uk}^{(T)}), \\ \hat{m}_{uk}^{(T)} &\sim \text{Poisson}\Big[-\frac{\phi_{uk}^{(T-1)}}{\tau} \ln(1 - \varrho_{uk}^{(T)})\Big]. \end{aligned} \tag{3.2.18}$$

where

$$\varrho_{uk}^{(t)} = \frac{\omega_{uk}^{(t)} - \frac{1}{\tau}\ln(1 - \varrho_{uk}^{(t+1)})}{\tau + \omega_{uk}^{(t)} - \frac{1}{\tau}\ln(1 - \varrho_{uk}^{(t+1)})}. \tag{3.2.19}$$

Given $\tilde{A}^{(T-1)}_{uk\cdot\cdot} \sim \text{Poisson}(\phi^{(T-1)}_{uk}\omega^{(T-1)}_{uk})$, via the Poisson additive property, we have

$$\hat{m}^{(T)}_{uk} + \tilde{A}^{(T-1)}_{uk\cdot\cdot} \sim \text{Poisson}\Big(\phi^{(T-1)}_{uk}\Big[\omega^{(T-1)}_{uk} - \frac{1}{\tau}\ln(1-\varrho^{(T)}_{uk})\Big]\Big). \tag{3.2.20}$$

Combining the likelihood in Eq. (3.2.20) with the gamma prior placed on $\phi^{(T-1)}_{uk}$, we immediately have the conditional distribution of $\phi^{(T-1)}_{uk}$ via the gamma-Poisson conjugacy as

$$\begin{aligned}(\phi^{(T-1)}_{uk}|-) \sim \text{Gamma}\Big[&\phi^{(T-2)}_{uk}/\tau + \hat{m}^{(T)}_{uk} + \tilde{A}^{(T-1)}_{uk\cdot\cdot}, \\ &\frac{1}{\tau + \omega^{(T-1)}_{uk} - \frac{1}{\tau}\ln(1-\varrho^{(T)}_{uk})}\Big].\end{aligned} \tag{3.2.21}$$

Here, $\hat{m}^{(T)}_{uk}$ can be considered as the *backward* information passed from $t = T$ to $T-1$. Recursively, we augment $\phi^{(t)}_{uk}$ at each time slice with an auxiliary variable $\hat{m}^{(t)}_{uk}$ as

$$\begin{aligned}\hat{m}^{(t+1)}_{uk} + \tilde{A}^{(t)}_{uk\cdot\cdot} &\sim \text{NB}(\phi^{(t-1)}_{uk}/\tau, \varrho^{(t)}_{uk}), \\ \hat{m}^{(t)}_{uk} &\sim \text{CRT}(\tilde{A}^{(t)}_{uk\cdot\cdot} + \hat{m}^{(t+1)}_{uk}, \phi^{(t-1)}_{uk}/\tau),\end{aligned} \tag{3.2.22}$$

where the NB distribution over $\hat{m}^{(t+1)}_{uk} + \tilde{A}^{(t)}_{uk\cdot\cdot}$ is obtained via the Poisson additive property and gamma-Poisson conjugacy with $\tilde{A}^{(t)}_{uk\cdot\cdot} \sim \text{Poisson}(\phi^{(t)}_{uk}\omega^{(t)}_{uk})$. Repeatedly exploiting the Poisson logarithmic bivariate distribution, we have

$$\begin{aligned}\hat{m}^{(t+1)}_{uk} + \tilde{A}^{(t)}_{uk\cdot\cdot} &\sim \text{SumLog}(\hat{m}^{(t)}_{uk}, \varrho^{(t)}_{uk}), \\ \hat{m}^{(t)}_{uk} &\sim \text{Poisson}\Big[-\frac{\phi^{(t-1)}_{uk}}{\tau}\ln(1-\varrho^{(t)}_{uk})\Big].\end{aligned}$$

By repeatedly exploiting the Poisson additive property and gamma-Poisson conjugacy, we obtain

$$\begin{aligned}(\phi^{(t-1)}_{uk} \mid -) \sim \text{Gamma}\Big[&\hat{m}^{(t)}_{uk} + \phi^{(t-2)}_{uk}/\tau + \tilde{A}^{(t-1)}_{uk\cdot\cdot}, \\ &\frac{1}{\tau + \omega^{(t-1)}_{uk} - \frac{1}{\tau}\ln(1-\varrho^{(t)}_{uk})}\Big].\end{aligned} \tag{3.2.23}$$

We sample the auxiliary variables $\hat{m}^{(t)}_{uk}$ and update $\varrho^{(t)}_{uk}$ recursively from $t = T$ to $t = 1$, which can be considered as the *backward filtering* step. Then, in the *forward* pass we sample $\phi^{(t)}_{uk}$ from $t = 1$ to $t = T$.

Sampling hyperparameters. Via the gamma-gamma conjugacy, we sample c_0 and β as

$$\begin{aligned}(c_0 \mid -) &\sim \text{Gamma}\Big[0.1 + \gamma_0, 1/(0.1 + \sum_k r_k)\Big], \\ (\beta \mid -) &\sim \text{Gamma}\Big[0.1 + \sum_{k,k'} r_k \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}}, \frac{1}{0.1 + \sum_{k,k'}\Omega_{kk'}}\Big].\end{aligned} \tag{3.2.24}$$

Algorithm 1 summarizes the full procedure.

Time complexity of parameter estimation: For the proposed DPGM, sampling $\{\tilde{A}^{(t)}_{uv}\}_{u,v,t}$ and $\{\tilde{A}^{(t)}_{ukk'v}\}_{u,v,k,k',t}$ takes $\mathcal{O}(N^e\bar{K}^2)$ with N^e being the number of non-zero entries. Sampling $\{\phi^{(t)}_{uk}\}_{i,k,t}$ takes $\mathcal{O}(V\bar{K}T)$ and sampling $\{\Omega_{kk'}\}_{k,k',t}$ takes $\mathcal{O}(\bar{K}^2)$. Overall, the computational complexity of DPGM is $\mathcal{O}(N^e\bar{K}^2 + V\bar{K}T + \bar{K}^2)$.

Algorithm 1 Gibbs sampling algorithm for the dynamic Poisson gamma memberships model (DPGM)

```
Input: relational data $\{A^{(t)}\}_{t=1}^{T}$, iterations $\mathcal{J}$.
Initialize the maximum number of communities $K$, hyperparameters $\gamma_0, \beta, c, \tau$.
for $iter = 1$ to $\mathcal{J}$ do
    Sample $\{\tilde{A}_{uv}^{(t)}\}_{u,v,t}$ for non-zero edges (Eq. 3.2.6)
    Sample $\{\tilde{A}_{ukk'v}^{(t)}\}_{u,v,k,k',t}$ (Eq. 3.2.7) and update
        $\tilde{A}_{\cdot kk'\cdot}^{(\cdot)} = \sum_t \sum_{u,v\neq u} \tilde{A}_{ukk'v}^{(t)}$
        $\tilde{A}_{uk\cdot\cdot}^{(t)} = \sum_{v\neq u,k'} \tilde{A}_{ukk'j}^{(t)}$
    for $t = T$ to $1$ do
        Sample $\{\hat{m}_{uk}^{(t)}\}_{u,k}$ (Eq. 3.2.22) and update $\{\varrho_{uk}^{(t)}\}_{u,k}$ (Eq. 3.2.19)
    end for
    for $t = 1$ to $T$ do
        Sample $\{\phi_{uk}^{(t)}\}_{u,k}$ (Eq. 3.2.23)
    end for
    Sample $\{l_{kk'}\}_{k,k'}$ (Eq. 3.2.9) and update
        $\theta_{kk'} = \sum_{t,u,v\neq u} \phi_{uk}^{(t)} \phi_{vk'}^{(t)}, \quad \tilde{p}_{kk'} = \frac{\rho_{kk'}}{\rho_{kk'}+\beta}$
    Sample $\{\Omega_{kk'}\}_{k,k'}$ (Eq. 3.2.13), $\{r_k\}_k$ (Eq. 3.2.11), and $\xi$ (Eq. 3.2.12)
end for
Output posterior means: $\{\phi_{uk}^{(1:T)}\}_{u,k}, \{r_k\}_k, \xi, \{\Omega_{kk'}\}_{k,k'}$.
```

3.2.4 *Online Gibbs Sampling*

To accommodate the proposed model with large-scale dynamic relational data, we propose an online Gibbs sampling algorithm based on the recent developed Bayesian conditional density filtering (BCDF) (Guhaniyogi et al. 2014), which has been adapted for knowledge graph learning (Hu et al. 2016b) and Poisson tensor factorization (Hu et al. 2015) recently. The main idea of BCDF is to partition the data into small mini-batches, and then to perform inference by updating the sufficient statistics using each mini-batch in each iteration. Specifically, the sufficient statistics used in our model are the latent count numbers. The main procedure of our online Gibbs sampler is: We use Z and Z^i to denote the indices of the entire data and the mini-batch in ith iteration respectively. We define the quantities updated with the mini-batch in ith iteration as:

$$\tilde{A}_{uk\cdot\cdot}^{(t)i} = \frac{|Z|}{|Z^i|} \sum_{\substack{v\neq u,\\ u,v\in Z_t}} \sum_{k'} \tilde{A}_{ukk'v'}^{(t)}$$

$$\tilde{A}_{\cdot kk'\cdot}^{i} = \frac{|Z|}{|Z^i|} \sum_{t} \sum_{\substack{u,v\neq u,\\ u,v\in Z_t}} \tilde{A}_{ukk'v}^{(t)}.$$

Then, we update the sufficient statistics used to sample model parameters as

$$\begin{aligned}
\tilde{A}^{(t)i}_{uk\cdot\cdot} &= (1-\rho^i)\tilde{A}^{i-1}_{uk\cdot\cdot} + \rho^i \frac{|Z|}{|Z^i|} \sum_{\substack{v\neq u,\\ u,v\in Z_t}} \sum_{k'} \tilde{A}^{(t)}_{ukk'v'} \\
\tilde{A}^{i}_{\cdot kk'\cdot} &= (1-\rho^i)\tilde{A}^{i}_{\cdot kk'\cdot} + \rho^i \frac{|Z|}{|Z^i|} \sum_{t} \sum_{\substack{u,v\neq u,\\ u,v\in Z_t}} \tilde{A}^{(t)}_{ukk'v'}
\end{aligned}$$

where $\rho^i = (i+i_0)^{-\kappa}$, where $i_0 > 0$ and $\kappa \in (1/2, 1]$, is the decaying factor. In the online Gibbs sampling, we calculate the sufficient statistics with each mini-batch data and resample the model parameters and hyperparameters using the procedure in batch Gibbs sampling algorithm.

3.3 THE DEPENDENT RELATIONAL GAMMA PROCESS MODEL

In the previous section, we have introduced the Dynamic Poisson gamma memberships model that only captures the evolving node behaviours but assumes the underlying latent communites are static over time. Nevertheless, these latent communities composed of nodes can also form and decay over time in real-world networks. Hence, we will exploit the time-dependent relational gamma process to capture the birth and death dynamics of latent communities in this section.

3.3.1 *Model of active communities.*

Many previous works (M. Kim et al. 2013; K. Xu 2015) have shown that explicitly modelling the dynamics of latent communities using a distance-dependent Indian buffet process (dd-IBP) (Gershman et al. 2015) or a linear dynamical system discovers interpretable latent structure, and thus achieves good predictive performance. Here we build the community interaction weight $\Omega_{kk'}$ on the relational gamma process construction (M. Zhou et al. 2015). That is, we first generate a community weight r_k independently for each community k as $r_k \sim \text{Gamma}(\gamma_0/K, 1/c)$, where γ_0 denotes the concentration parameter, $1/c$ denotes the scale parameter, and K is the maximum number of communities. Then, the inter-community interaction weight $\Omega_{kk'}$ and intra-community interaction weight Ω_{kk} can be generated as

$$\Omega_{kk'} \sim \begin{cases} \text{Gamma}(\xi r_k, 1/\beta), & \text{if } k = k' \\ \text{Gamma}(r_k r_{k'}, 1/\beta), & \text{otherwise} \end{cases} \tag{3.3.1}$$

where $\xi \sim \text{Gamma}(1,1)$ and $\beta \in \mathbb{R}_{>0}$. For dynamic relational data, we exploit the thinned completely random measures framework to capture the birth/death dynamics of latent communities assuming that the status of community k can be either *active* or *inactive* at time t. More specifically, we use a Bernoulli random variable $b_k^{(t)} = 1$ to indicate the presence of community k at time t, and $b_k^{(t)} = 0$ otherwise. Accordingly, the interaction weight between community k and k' is active at time t only if the two communities are both active at that time: $\Omega^{(t)}_{kk'} = \Omega_{kk'} b_k^{(t)} b_{k'}^{(t)}$.

Given the community interaction weight matrix Ω defined in Eq. (3.3.1), we generate the time-dependent community interaction weights $\Omega_{kk'}^{(t)}$ using the thinning function introduced in Section 2.2.10 with the prior specification:

$$\tilde{\omega}_{lk} \sim \mathcal{NIG}(0,1,1), \;\; \tilde{\theta}_k \sim \text{Categorical}(\tilde{\theta}_1,\ldots,\tilde{\theta}_D),$$

$$P_{x_k}(t) = \sigma\left\{\tilde{\omega}_{0k} + \sum_{l=1}^{T} \tilde{\omega}_{lk} \exp[-\tilde{\theta}_k (t-l)^2]\right\},$$

where we fix the centres of the covariate-dependent kernel functions to the T discrete time points of the considered dynamic network. The probability of activity/inactivity of community k at time t can be determined by the thinning function. A smooth thinning function can encourage the snapshots of a dynamic network at nearby covariate values t to share a similar set of communities.

The full generative model for the observed dynamic network data $\{A^{(t)}\}_{t\in\mathcal{T}}$ along with the latent variables, parameters, and hyperparameters, is given by

$$\begin{aligned}
r_k &\sim \text{Gamma}(\gamma_0/K, 1/c), \qquad (3.3.2)\\
\theta_{uk} &\sim \text{Gamma}(1,1),\\
\xi &\sim \text{Gamma}(1,1),\\
\Omega_{kk'} &\sim \begin{cases} \text{Gamma}(\xi r_k, 1/\beta), & \text{if } k = k' \\ \text{Gamma}(r_k r_{k'}, 1/\beta), & \text{otherwise} \end{cases}\\
\tilde{\omega}_{lk} &\sim \mathcal{NIG}(0,1,1),\\
\tilde{\theta}_k &\sim \text{Categorical}(\tilde{\theta}_1,\ldots,\tilde{\theta}_D),\\
b_k^{(t)} &\sim \text{Bernoulli}\left(\sigma\left\{\tilde{\omega}_{0k} + \sum_{l=1}^{T} \tilde{\omega}_{lk} \exp[-\tilde{\theta}_k (t-l)^2]\right\}\right),\\
\phi_{uk}^{(t)} &\sim \text{Gamma}(\phi_{uk}^{(t-1)}/\tau, 1/\tau),\\
\phi_{uk}^{(1)} &\sim \text{Gamma}(\theta_{uk}/\tau, 1/\tau),\\
\Omega_{kk'}^{(t)} &= b_k^{(t)} \Omega_{kk'} b_{k'}^{(t)},\\
\tilde{A}_{uv}^{(t)} &\sim \text{Poisson}\left(\sum_{k,k'=1}^{K} \phi_{uk}^{(t)} \Omega_{kk'}^{(t)} \phi_{vk'}^{(t)}\right),\\
A_{uv}^{(t)} &= \mathbb{1}(\tilde{A}_{uv}^{(t)} \geq 1).
\end{aligned}$$

Note that dynamic networks characterized by both homophily and stochastic equivalence (P. Hoff 2008) can be appropriately modelled via the proposed framework in Eq. (3.3.2) as discussed in (M. Zhou et al. 2015).

3.3.2 *Bayesian nonparametric interpretation.*

As $K \to \infty$, the community weights and their corresponding node membership parameters constitute a draw from a gamma process as $G = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$, where $\theta_k = (\theta_{1k},\ldots,\theta_{uk}) \in \Theta$ is an atom sampled from a N-dimensional base measure $G_0(\mathrm{d}\theta_k)/G_0(\Theta) = \prod_{u=1}^{N} \text{Gamma}(1,1)$. Accordingly, the intra- and inter-community interaction weights and their corresponding pair of scale parameters constitute a draw $\Omega \mid G = \sum_{k=1}^{\infty}\sum_{k'=1}^{\infty} \Omega_{kk'} \delta_{(\theta_k,\theta_{k'})}$ from a relational gamma process (M.

Zhou et al. 2015). Via the thinned CRMs framework, $\Omega^{(t)} \mid \Omega = \sum_{k=1}^{\infty} \sum_{k'=1}^{\infty} b_k^{(t)} b_{k'}^{(t)} \Omega_{kk'} \delta_{(\theta_k, \theta_{k'})}$ can be viewed as a draw from a covariate-dependent relational gamma process.

3.3.3 *Inference algorithm*

An efficient Gibbs sampling algorithm is derived for the inference of parameters and hyperparameters of interest in the proposed model. Let $A^{(1:t)}$ denotes the sequence $A^{(1)}, \ldots, A^{(t)}$ and similarly for $\Phi^{(1:t)}$ and $\Omega^{(1:t)}$. The model parameters that need to be sampled include: latent node-community memberships $\{\phi_{uk}^{(t)}\}$, $\{\theta_{uk}\}$, individual community weights $\{r_k\}$, scale parameter ξ, communities interaction weights $\{\Omega_{kk'}\}$, kernel weights $\{\tilde{\omega}_{lk}\}$, kernel widths $\{\tilde{\theta}_k\}$, thinning variables $\{b_k^{(t)}\}$, and latent counts $\{\tilde{A}_{uv}^{(t)}\}$. Exploiting the Pólya-gamma data augmentation technique (Polson et al. 2013) and the data augmentation and marginalization technique (M. Zhou and L. Carin 2015b), a simple and efficient Gibbs sampling algorithm is developed to perform the model inference.

Sampling latent count $\tilde{A}_{uv}^{(t)}$: We sample the latent count $\tilde{A}_{uv}^{(t)}$ as

$$(\tilde{A}_{uv}^{(t)} \mid -) \sim A_{uv}^{(t)} \mathrm{Poisson}_{+} \left(\sum_{k=1}^{K} \sum_{k'=1}^{K} \Omega_{kk'}^{(t)} \phi_{uk}^{(t)} \phi_{vk'}^{(t)} \right). \tag{3.3.3}$$

Sampling latent subcount $\tilde{A}_{ukk'v}^{(t)}$: To update the node-community memberships $\{\phi_{uk}^{(t)}\}_{u,k,t}$ and community-community interaction weights $\{\Omega_{kk'}^{(t)}\}_{k,k',t}$, we need to partition the count $\tilde{A}_{uv}^{(t)}$ into the sub counts $\{\tilde{A}_{ukk'v}^{(t)}\}_{k,k'}$, where $\tilde{A}_{ukk'v}^{(t)}$ measures the interaction strength between nodes u and v due to their associations to communities k and k', respectively. Via the Poisson-multinomial equivalence, we sample the latent subcounts $\tilde{A}_{ukk'v}^{(t)}$ as

$$(\tilde{A}_{ukk'v}^{(t)} \mid -) \sim \mathrm{Multinomial} \left(\tilde{A}_{uv}^{(t)}; \frac{\Omega_{kk'}^{(t)} \phi_{uk}^{(t)} \phi_{vk'}^{(t)}}{\sum_{k=1}^{K} \sum_{k'=1}^{K} \Omega_{kk'}^{(t)} \phi_{uk}^{(t)} \phi_{vk'}^{(t)}} \right). \tag{3.3.4}$$

Sampling node-community memberships $\Phi^{(0:T)}$: We can exploit the Poisson logarithmic bivariate distribution to recursively sample the auxiliary variables $\hat{m}_{uk}^{(t)}$ and update $\eta_{uk}^{(t)}$ backwardly from $t = T$ to 1 as we did in deriving the Gibbs sampler for the dynamic Poisson gamma memberships model:

$$\hat{m}_{uk}^{(t)} \sim \mathrm{CRT}(\hat{m}_{uk}^{(t+1)} + \tilde{A}_{uk\cdot\cdot}^{(t)}, \phi_{uk}^{(t-1)}/\tau), \tag{3.3.5}$$

$$\eta_{uk}^{(t)} = \frac{\psi_{uk}^{(t)} - \frac{1}{\tau}\ln(1 - \eta_{uk}^{(t+1)})}{\tau + \psi_{uk}^{(t)} - \frac{1}{\tau}\ln(1 - \eta_{uk}^{(t+1)})}, \tag{3.3.6}$$

where we define $\phi_{uk}^{(0)} = \theta_{uk}$, $\hat{m}_{uk}^{(T+1)} = 0$, and $\eta_{uk}^{(T+1)} = 0$. We then sample the $\phi_{uk}^{(t)}$ forwardly from $t = 0$ to T as

$$(\theta_{uk} \mid -) \sim \mathrm{Gamma}\Big[1 + \hat{m}_{uk}^{(1)}, \frac{1}{1 - \frac{1}{\tau}\ln(1 - \eta_{uk}^{(1)})}\Big], \tag{3.3.7}$$

$$(\phi_{uk}^{(t)} \mid -) \sim \mathrm{Gamma}\Big[\hat{m}_{uk}^{(t+1)} + \phi_{uk}^{(t-1)} + \tilde{A}_{uk\cdot\cdot}^{(t)}, \frac{1}{\tau + \psi_{uk}^{(t)} - \frac{1}{\tau}\ln(1 - \eta_{uk}^{(t)})}\Big], \quad \text{for } t \in \mathcal{T}. \tag{3.3.8}$$

Marginalization over Ω, r: We define the latent Poisson count as

$$\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} \equiv 2^{-\delta_{kk'}} \sum_t \sum_u \sum_{v \neq u} \tilde{A}^{(t)}_{ukk'v'}$$

where $\delta_{kk'} = 1$ if $k = k'$, and $\delta_{kk'} = 0$ otherwise. Via the Poisson additive property, we have

$$\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} \sim \text{Poisson}(\Omega_{kk'}\rho_{kk'}),$$

where $\rho_{kk'} \equiv \sum_t b_k^{(t)} b_{k'}^{(t)} \sum_u \sum_{v\neq u} \phi^{(t)}_{uk} \phi^{(t)}_{vk'}$.

As we have the prior specification $\Omega_{kk'} \sim \text{Gamma}(r_k \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}}, 1/\beta)$, marginalizing over $\Omega_{kk'}$ yields

$$\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} \sim \text{NB}(r_k \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}}, \chi_{kk'}),$$

where $\chi_{kk'} \equiv \frac{\rho_{kk'}}{\beta + \rho_{kk'}}$.

To marginalize over r_k, we introduce an auxiliary variable:

$$l_{kk'} \sim \text{CRT}(\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}, r_k \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}}), \tag{3.3.9}$$

and then re-express the joint distribution over $\tilde{A}^{(\cdot)}_{\cdot kk'\cdot}$ and $l_{kk'}$ as

$$\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} \sim \text{SumLog}(l_{kk'}, \chi_{kk'}), \quad l_{kk'} \sim \text{Poisson}[-r_k \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}} \ln(1 - \chi_{kk'})].$$

Via the Poisson additive property, we have

$$l_{k\cdot} \equiv \sum_{k'} l_{kk'} \sim \text{Poisson}[-r_k \sum_{k'} \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}} \ln(1 - \chi_{kk'})].$$

Sampling community interaction weights Ω: Via the gamma Poisson conjugacy, we sample $\Omega_{kk'}$ from its conditional posterior as

$$(\Omega_{kk'} \mid -) \sim \text{Gamma}\left[\tilde{A}^{(\cdot)}_{\cdot kk'\cdot} + r_k \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}}, \frac{1}{\beta + \rho_{kk'}}\right]. \tag{3.3.10}$$

Sampling community weight r_k: Using the gamma-Poisson conjugacy, we sample r_k as

$$(r_k \mid -) \sim \text{Gamma}\left[\frac{\gamma_0}{K} + \sum_{k'} l_{kk'}, \frac{1}{c - \sum_{k'} \xi^{\delta_{kk'}} r_{k'}^{1-\delta_{kk'}} \ln(1 - \chi_{kk'})}\right]. \tag{3.3.11}$$

Sampling ξ: Using the gamma-Poisson conjugacy, we sample

$$(\xi \mid -) \sim \text{Gamma}\left[1 + \sum_k l_{kk}, \frac{1}{1 - \sum_k r_k \ln(1 - \chi_{kk})}\right]. \tag{3.3.12}$$

Sampling thinning variable $b_k^{(t)}$: If $\sum_u \tilde{A}^{(t)}_{uk\cdot\cdot} > 0$, we set $b_k^{(t)} = 1$, and if $\sum_u \tilde{A}^{(t)}_{uk\cdot\cdot} = 0$, we sample $b_k^{(t)}$ by the following process: we define fictitious latent counts $\varpi_k^{(t)} \sim \text{Poisson}(r_k \xi \rho_{kk})$ disregarding $b_k^{(t)}$ to determine whether $\sum_u \tilde{A}^{(t)}_{uk\cdot\cdot} = 0$ because community k has been thinned or because community k has not been observed at time t. We thus sample $b_k^{(t)}$ when $\sum_u \tilde{A}^{(t)}_{uk\cdot\cdot} = 0$ as

1. If $\varpi_k^{(t)} = 0$, we sample $b_k^{(t)}$ as

$$\begin{aligned} p(b_k^{(t)} = 1 \mid \varpi_{kk}^{(t)} = 0) &\propto p(b_k^{(t)} = 1)\mathrm{Poisson}(0; r_k \xi \rho_{kk}), \\ p(b_k^{(t)} = 0 \mid \varpi_{kk}^{(t)} = 0) &\propto p(b_k^{(t)} = 0)\mathrm{Poisson}(0; r_k \xi \rho_{kk}). \end{aligned} \tag{3.3.13}$$

2. If $\varpi_k^{(t)} > 0$, we sample $b_k^{(t)}$ as

$$p(b_k^{(t)} = 1 \mid \varpi_{kk}^{(t)} > 0) \propto p(b_k^{(t)} = 1)\left[1 - \mathrm{Poisson}(0; r_k \xi \rho_{kk})\right]. \tag{3.3.14}$$

Sampling kernel weights $\tilde{\omega}$: The normal-inverse-gamma prior placed over $\tilde{\omega}_{lk}$ can be equivalently generated from the following process by introducing auxiliary variables $\{\vartheta_{lk}\}$:

$$\vartheta_{lk} \sim \mathrm{Gamma}(1,1), \quad \tilde{\omega}_k \sim \mathcal{N}(0, \Sigma_\vartheta),$$

where $\tilde{\omega}_k = (\tilde{\omega}_{1k}, \ldots, \tilde{\omega}_{Lk})$ and $\Sigma_\vartheta = \mathrm{diag}(\vartheta_{0k}, \ldots, \vartheta_{Lk})$.
Let $\mathcal{K}_{tk} = (1, \mathcal{K}(t, t_1, \tilde{\theta}_k), \ldots, \mathcal{K}(t, t_L, \tilde{\theta}_k))^{\mathrm{T}}$ be the vector of the kernels evaluated at time t. We sample $\{\tilde{\omega}_{lk}\}$ exploiting the Pólya-gamma data augmentation technique (Polson et al. 2013) for logistic regression by introducing auxiliary variables as

$$(\tilde{b}_{kt} \mid -) \sim \mathrm{PG}(1, \mathcal{K}_{tk}^{\mathrm{T}} \tilde{\omega}_k),$$

where $\mathrm{PG}(a, b)$ denotes the Pólya-gamma distribution with $a \in \mathbb{R}$ and $b > 0$. Let $\tilde{\Omega}(\tilde{b}_k)$ denote the $T \times T$ diagonal matrix whose t-th diagonal element is $\tilde{b}_{kt}$, and let $\mu_k = (b_k^{(1)} - 1/2, \ldots, b_k^{(T)} - 1/2)^{\mathrm{T}}$. The conditional distribution of $\tilde{\omega}_k$ is

$$(\tilde{\omega}_k \mid -) \sim \mathcal{N}(\mu_{\tilde{\omega}_k}, \Sigma_{\tilde{\omega}_k}), \tag{3.3.15}$$

where $\Sigma_{\tilde{\omega}_k} = (\Sigma_\vartheta^{-1} + \mathcal{K}_{tk}^{\mathrm{T}} \tilde{\Omega}(\tilde{b}_k) \mathcal{K}_{tk})^{-1}$ and $\mu_{\tilde{\omega}_k} = \Sigma_{\tilde{\omega}_k} \mathcal{K}_{tk}^{\mathrm{T}} \mu_k$.

We sample ϑ_{lk} from its conditional posterior via the gamma normal conjugacy as

$$(\vartheta_{lk} \mid -) \sim \mathrm{Gamma}\Big(\frac{3}{2}, \frac{1}{1 + \frac{1}{2}\tilde{\omega}_{lk}^2}\Big).$$

Sampling kernel width $\tilde{\theta}$: We uniformly draw $\tilde{\theta}_k$ from a fixed dictionary $\{\tilde{\theta}_1^*, \ldots, \tilde{\theta}_D^*\}$ of size D, and sample $\tilde{\theta}_k$ as

$$p(\tilde{\theta}_k = \tilde{\theta}_d^* \mid -) \propto \frac{1}{D} \prod_{t \in \mathcal{T}} \left(P_{\tilde{\theta}_d^*}(t)\right)^{b_k^{(t)}} \left(1 - P_{\tilde{\theta}_d^*}(t)\right)^{1 - b_k^{(t)}} \tag{3.3.16}$$

where the thinning function is denoted as a function of $\tilde{\theta}_d^*$ since the values of all the other variables are fixed as

$$P_{\tilde{\theta}_d^*}(t) = \sigma\Big\{\tilde{\omega}_{0k} + \sum_{l=1}^{T} \tilde{\omega}_{lk} \exp[-\tilde{\theta}_k (t - l)^2]\Big\}.$$

The full procedure of our Gibbs sampling algorithm is summarized in Algorithm 2.

Time complexity of parameter estimation: For the proposed DRGPM, sampling $\{\tilde{A}_{uv}^{(t)}\}_{u,v,t}$ and $\{\tilde{A}_{ukk'v}^{(t)}\}_{u,v,k,k',t}$ takes $\mathcal{O}(N^e \bar{K}^2)$ with N^e being the number of non-zero entries. Sampling $\{\phi_{uk}^{(t)}\}_{i,k,t}$ takes $\mathcal{O}(V\bar{K}T)$ and sampling $\{\Omega_{kk'}^{(t)}\}_{k,k',t}$ takes $\mathcal{O}(\bar{K}^2 T)$. Overall, the computational complexity of DRGPM is $\mathcal{O}(N^e \bar{K}^2 + V\bar{K}T + \bar{K}^2 T)$. The computational complexity of D-GPPF and DPGM is $\mathcal{O}(N^e \bar{K} + V\bar{K} + \bar{K}T)$ and $\mathcal{O}(N^e \bar{K}^2 + V\bar{K}T + \bar{K}^2)$, respectively.

Algorithm 2 Gibbs sampling algorithm for DRGPM

```
Input: dynamic relational data $\{A^{(t)}\}_{t=1}^{T}$, iterations $\mathcal{J}$.
Initialize the maximum number of communitys $K$, hyperparameters $\gamma_0, \beta, c$.
for $iter = 1$ to $\mathcal{J}$ do
    Sample $\{\tilde{A}_{uv}^{(t)}\}_{u,v}$ for non-zero edges (Eq. 3.3.3)
    Sample $\{\tilde{A}_{ukk'v}^{(t)}\}_{u,v,k,k'}$ (Eq. 3.3.4) and update
        $\tilde{A}_{\cdot kk' \cdot}^{(\cdot)} = \sum_t \sum_{u,v \neq u} \tilde{A}_{ukk'v}^{(t)}$
        $\tilde{A}_{uk\cdot\cdot}^{(t)} = \sum_{v \neq u,k'} \tilde{A}_{ukk'v}^{(t)}$
    Sample $\{l_{kk'}\}_{k,k'}$ (Eq. 3.3.9) and update
        $\rho_{kk'} = \sum_{t,u,v \neq u} b_k^{(t)} b_{k'}^{(t)} \phi_{uk}^{(t)} \phi_{vk'}^{(t)}, \quad \chi_{kk'} = \frac{\rho_{kk'}}{\rho_{kk'} + \beta}$
    Sample $\{r_k\}_k$ (Eq. 3.3.11), $\xi$ (Eq. 3.3.12), and $\{\Omega_{kk'}\}_{k,k'}$ (Eq. 3.3.10)
    for $t = T$ to 1 do
        Sample $\{\hat{m}_{uk}^{(t)}\}_{u,k}$ (Eq. 3.3.5) and update $\{\eta_{uk}^{(t)}\}_{u,k}$ (Eq. 3.3.6)
    end for
    for $t = 1$ to $T$ do
        Sample $\{\phi_{uk}^{(t)}\}_{u,k}$ (Eq. 3.3.8)
    end for
    Sample $\{\theta_{uk}\}_{u,k}$ (Eq. 3.3.7)
    for $t = 1$ to $T$ do
        Sample $\{b_k^{(t)}\}_k$ (Eqs. 3.3.13; 3.3.14)
    end for
    Sample $\{\tilde{\omega}_k\}_k$ (Eq. 3.3.15) and $\tilde{\boldsymbol{\theta}}$ (Eq. 3.3.16)
end for
Output posterior means: $\{\phi_{uk}^{(0:T)}\}_{u,k}, \{\theta_{uk}\}_{u,k}, \{r_k\}_k, \xi, \{\Omega_{kk'}\}_{k,k'}, \{b_k^{(1:T)}\}_k$.
```

3.4 EXPERIMENTS

In this section, we demonstrate the model capacity of the proposed dynamic Poisson gamma memberships model and the dynamic relational gamma process model on synthetic data. Quantitative evaluations compared with state-of-the-art methods are conducted on several real-world datasets in terms of missing link prediction and future network forecasting. The qualitative results on real-world military datasets demonstrate that the proposed dynamic network models discover the interpretable latent structure. To evaluate the performance of the proposed dynamic network models, the following state-of-the-art models are selected as baseline methods in the experiments:

1. **DRIFT:** the dynamic relational infinite feature model for which we used the code provided by (J. R. Foulds et al. 2011).[1]
2. **D-GPPF:** the dynamic gamma process Poisson factorization model, for which we set the hyperparameters and initialized the model parameters with the values provided in (Acharya et al. 2015b).
3. **DSBM:** the dynamic stochastic blockmodel (DSBM) based on an extended Kalman filter (EKF) augmented with a local search, for which we use the released code[2] with the default settings.
4. **HGPEPM:** the hierarchical gamma process edge partition model (HGPEPM) (M. Zhou et al. 2015)[3].

3.4.1 *Dynamic Community Detection*

First, we want to demonstrate how do the proposed dynamic network models detect time-evolving overlapping community structure using the nonnegative node-community memberships. We adapt the synthetic example used in Acharya et al. 2015b to generate a dynamic network of size 65 that evolves over five time slices as shown in column (a) of Figs. 3.2 to 3.4. More specifically, we generated three groups at $t = 1$, and split the second group at $t = 2$. From $t = 3$ to 4, the second and third group merge into one group. In Fig. 3.2, the discovered latent groups over time by D-GPPF are shown in column (b). D-GPPF captures the evolution of the discovered groups but has difficulties to characterize the changes of node-group memberships over time. As shown in column (b) of Fig. 3.3, the proposed model (DPGM) can detect the dynamic groups quite distinctively. We also show the associations of the nodes to the inferred latent groups by D-GPPF in column (c) of Fig. 3.2 and DPGM in column (c) of Fig. 3.3. In particular, we calculated the association weights of each node to the latent groups as $r_{tk}\phi_{uk}$ for D-GPPF and $\phi^t_{uk}\Omega_{kk}$ for DPGM. For both models, most of the redundant groups can effectively be shrunk even though we initialized both algorithms with $K = 50$. The node-group association weights estimated by DPGM vary smoothly over time and capture the evolution of the node-group memberships, which is consistent to the ground truth shown in column (a). In Fig. 3.4, column (b) and (c) show the inferred link probabilities and the node-community memberships, respectively. We found that DRGPM not only captures the evolving node-community structural changes but also effectively switches off most redundant communities.

---

[1] `http://jfoulds.informationsystems.umbc.edu/code/DRIFT.tar.gz`.

[2] `https://tinyurl.com/ydf29he9`.

[3] `https://github.com/mingyuanzhou/EPM`.

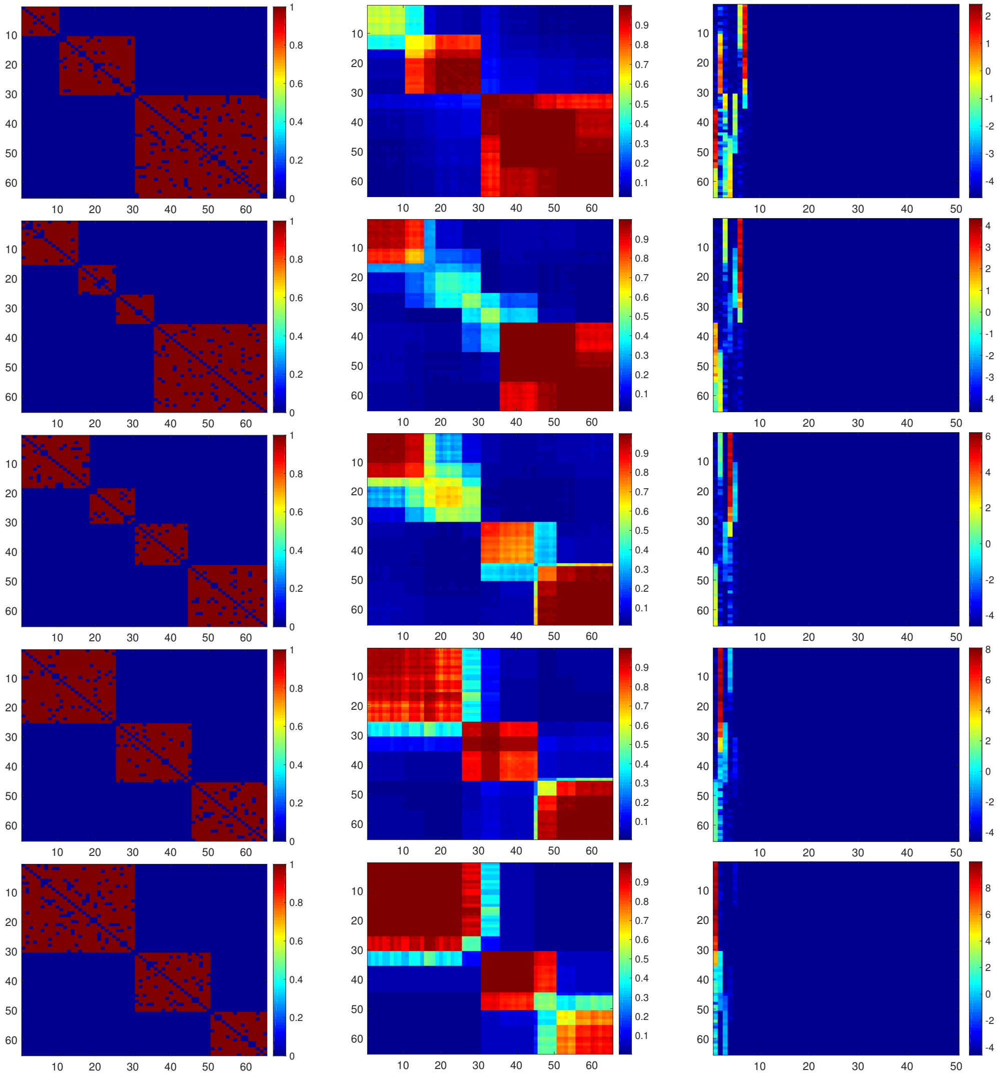
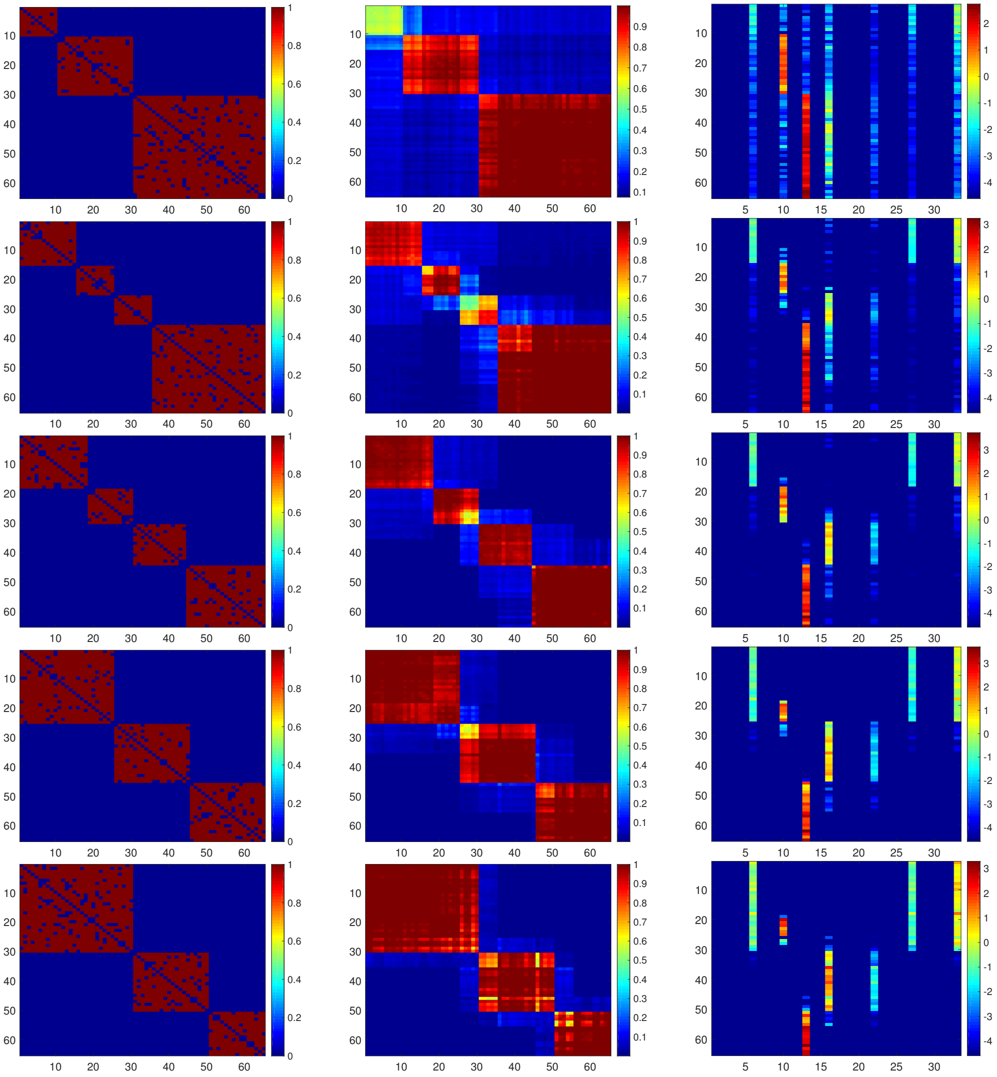
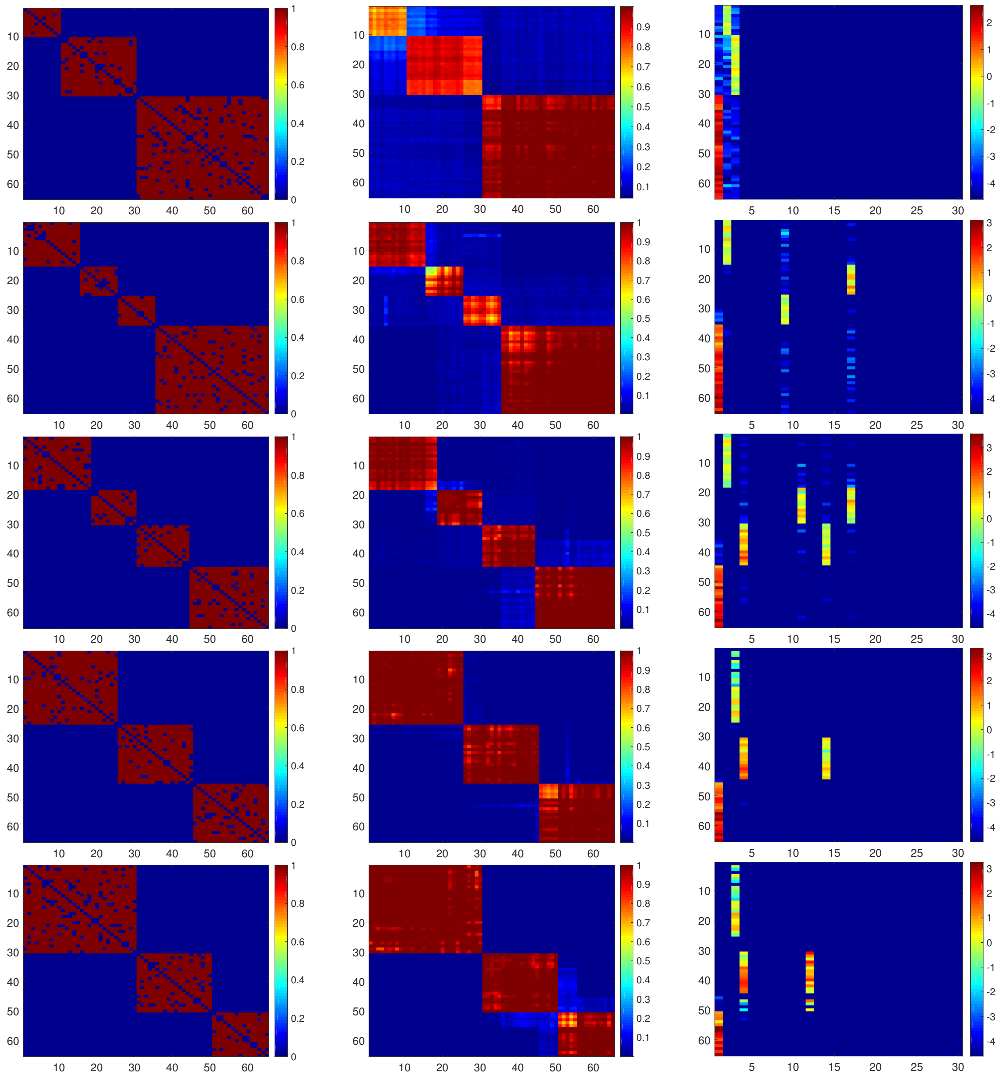


Figure 3.2: Dynamic community detection on synthetic data. We generated a dynamic network with five time snapshots as shown in column (a) ordered from top to bottom. The link probabilities estimated by D-GPPF are shown in column (b). The association weights of each node to the latent groups can be calculated by $r_{tk}\phi_{uk}$ for D-GPPF as shown in column (c). The pixel values are displayed on $\log_{10}$ scale.



(a) (b) (c)

Figure 3.3: Dynamic community detection on synthetic data. The simulated dynamic network with five time snapshots are shown in column (a) ordered from top to bottom. The link probabilities estimated by DPGM are shown in column (b). The association weights of each node to the latent groups can be calculated by $\phi_{uk}^{t}\Omega_{kk}$ for DPGM as shown in column (c). The pixel values are displayed on $\log_{10}$ scale.

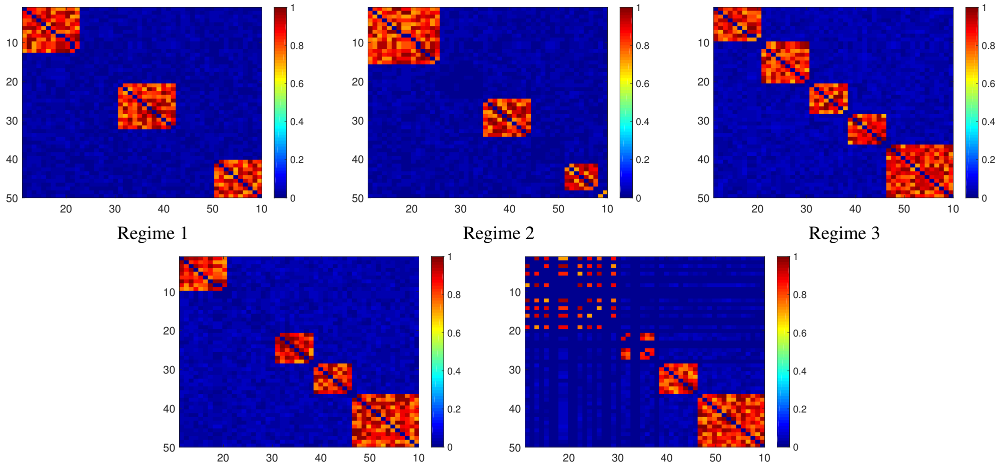


(a) (b) (c)

Figure 3.4: Dynamic community detection on synthetic data. The simulated dynamic network with five time snapshots are shown in column (a) ordered from top to bottom. The link probabilities estimated by DRGPM are shown in column (b). The association weights of each node to the latent groups can be calculated by $\phi_{uk}^{t}\Omega_{kk}^{t}$ for DRGPM as shown in column (c). The pixel values are displayed on $\log_{10}$ scale.

3.4.2 *Modelling Community Dynamics*

In the previous subsection, we have shown the dynamic overlapping community structure captured by the dynamic Poisson gamma memberships model (DPGM) and the dependent relational gamma process model (DRGPM). Next, we shall demonstrate how the dependent relational gamma process model characterizes the birth and death dynamics of latent communities over time. Following the procedure suggested by (Durante and Dunson 2016), we generated synthetic data to evaluate the proposed model in estimating the formation and evolution of the latent network sturcture. We consider a dynamic network with $N = 50$ nodes monitored for $T = 70$ equally spaced time snapshots. To generate a time-varying network, we first generated five regimes defining the true edge probabilities, as shown in Fig. 3.5. We then simulated the network edges $A_{uv}^{(t)} \mid \Pi_{uv}^{(t)} \sim \text{Bernoulli}(\Pi_{uv}^{(t)})$ with each of the five regimes according to Fig. 3.6. To demonstrate that DRGPM can infer interpretable latent structure while avoiding to overfit the data, we compared DRGPM with D-GPPF and DPGM. We initialized all methods setting $K = 30$.



Regime 1 Regime 2 Regime 3

Regime 4 Regime 5

Figure 3.5: True edge probabilities for the five regimes in the simulation study.

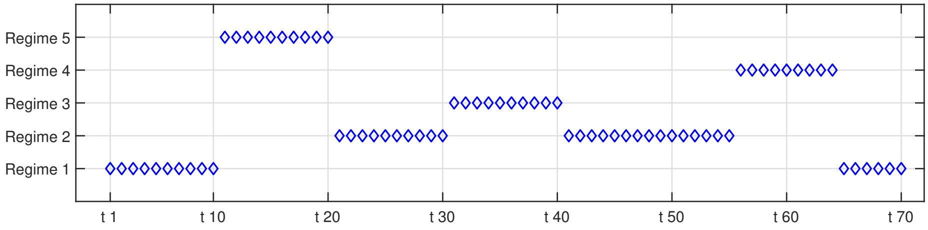
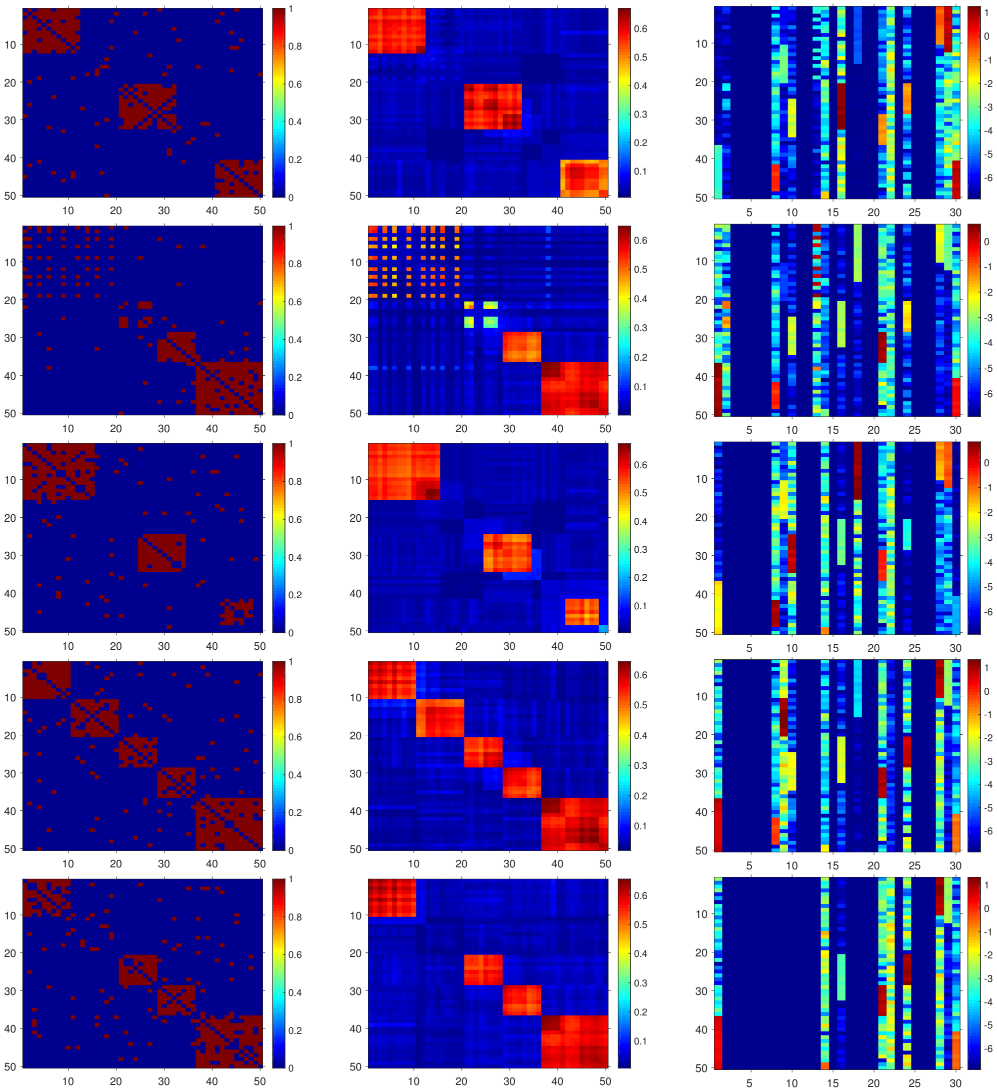


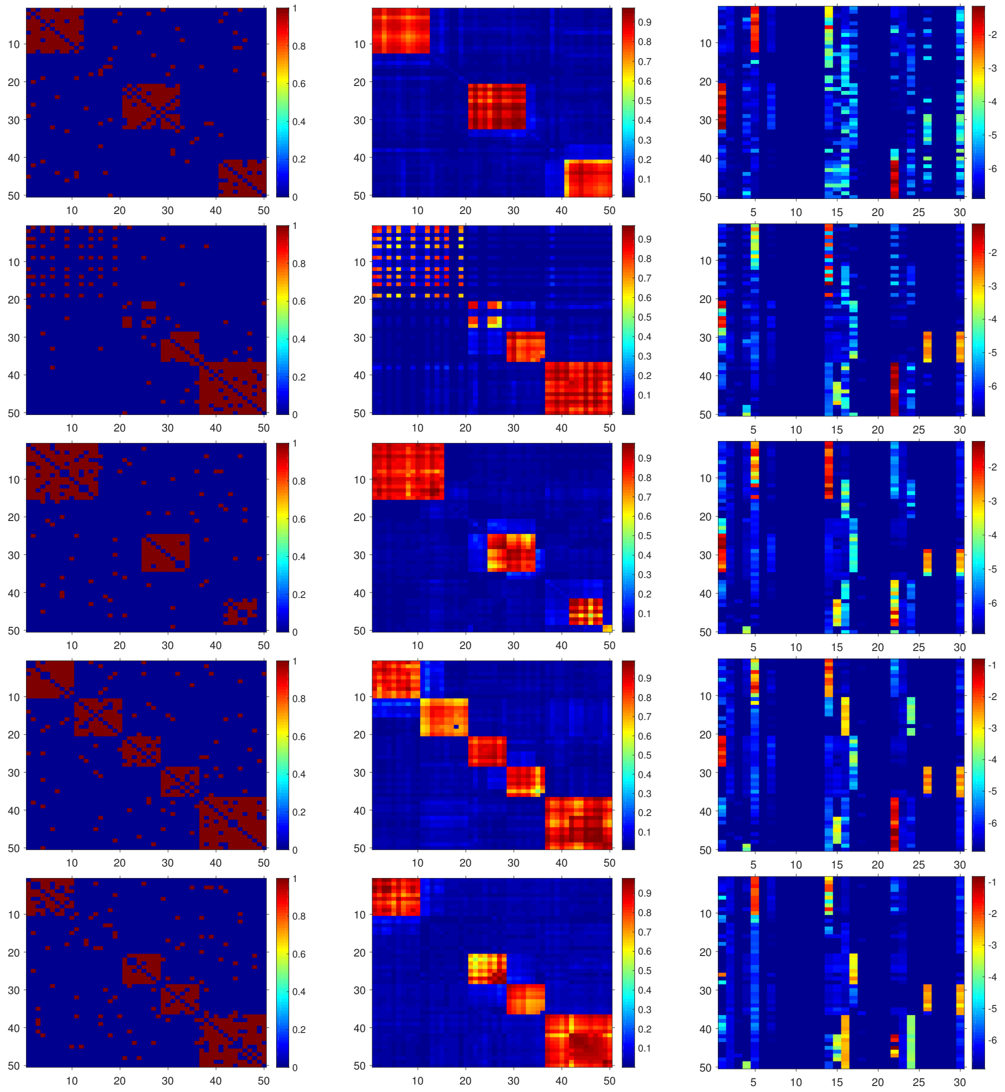
Figure 3.6: The graph showing which regime – i.e. true edge probabilities – for each snapshot is used to simulate the data.

In Fig. 3.7, the inferred link probabilities by D-GPPF are depicted in column (b). We also depicted the time-evolving node-community connections by computing the node-community association weights $r_k^{(t)} \phi_{uk}$ for D-GPPF in column (c). We note that D-GPPF needs to generate



(a) (b) (c)

Figure 3.7: We select five snapshots $(t = 5, 15, 25, 35, 60)$ of the simulated network as shown from top to bottom in column (a). The link probabilities inferred by D-GPPF are shown in column (b). The association weights of each node (row variable) to the groups (column variable), as shown in column (c), can be calculated as $r_k^{(t)}\phi_{uk}$ for D-GPPF. The pixel values are displayed on $\log_{10}$ scale.



(a) (b) (c)

Figure 3.8: The selected snapshots of the simulated network are shown in column (a). The link probabilities inferred by DPGM are shown in column (b). The association weights of each node (row variable) to the groups (column variable) are shown in column (c), can be calculated as $\phi_{uk}^{(t)}\Omega_{kk}$ for DPGM. The pixel values are displayed on $\log_{10}$ scale.

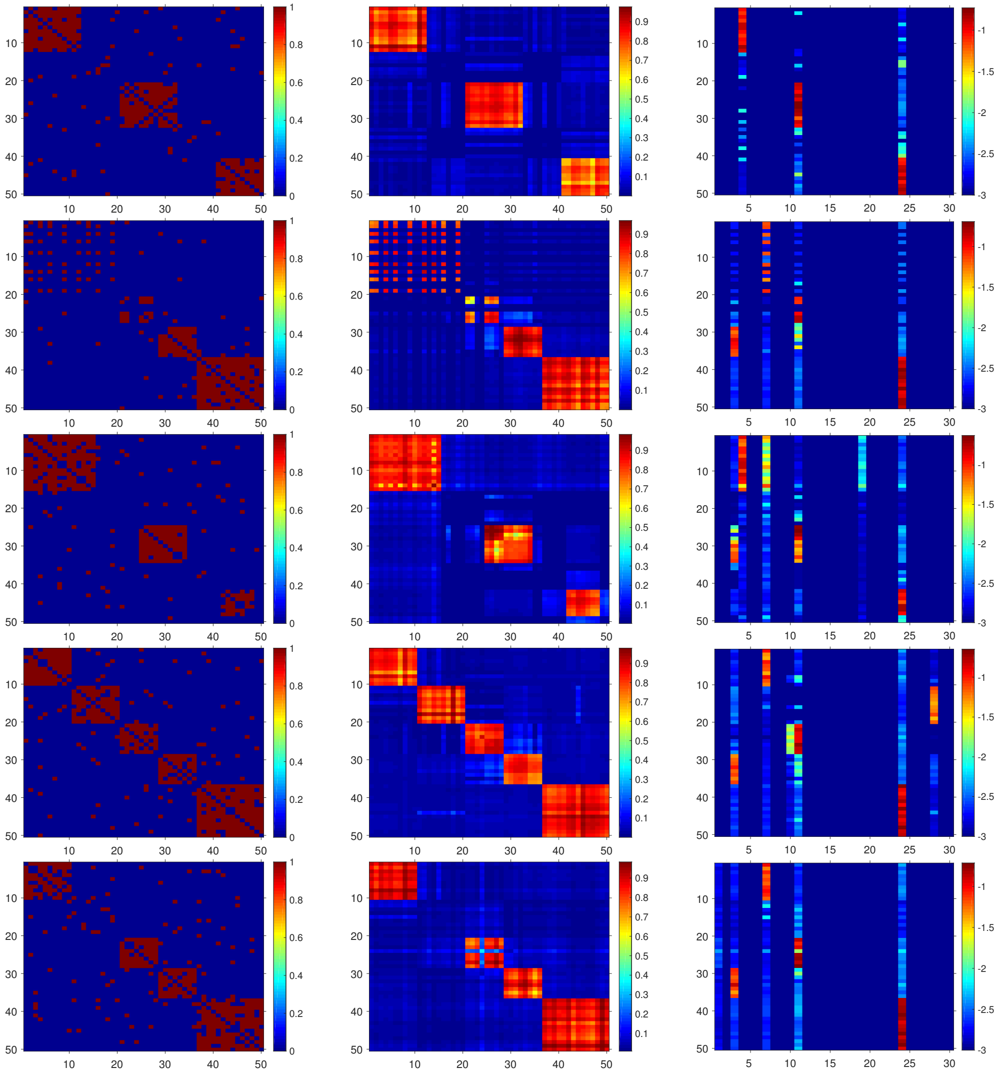
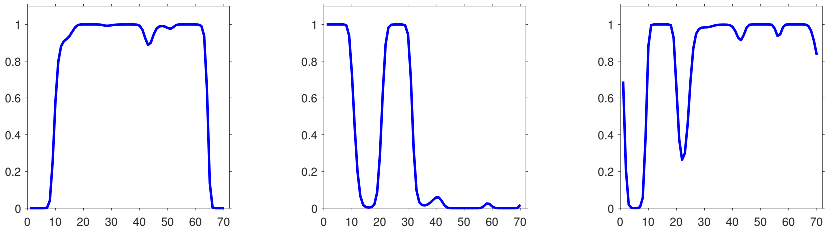
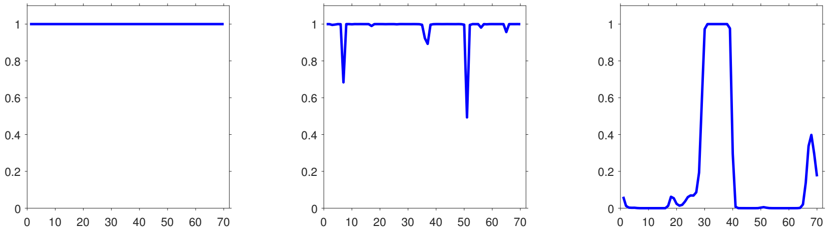


Figure 3.9: The selected snapshots of the simulated network are shown in column (a). The link probabilities inferred by DRGPM are shown in column (b). The association weights of each node (row variable) to the groups (column variable), as shown in column (c), can be calculated as $\phi_{uk}^{(t)}\Omega_{kk}^{(t)}$ for DRGPM. The pixel values are displayed on $\log_{10}$ scale.



Group 3 Group 4 Group 7



Group 11 Group 24 Group 28

Figure 3.10: The thinning probabilities (the mean of $b_k^{(t)}$) of the six active groups inferred by DRGPM.

many redundant communities to capture the time-evolving behavior of each node because of its unfavourable assumption that node memberships ϕ_{uk} are static while community weights $r_k^{(t)}$ are time-dependent. Without this restriction, DRGPM characterizes the evolving node-community associations by explicitly modelling time-dependent node-community memberships, and hence generates an appropriate number of communities as shown in columns (c) of Fig.3.9. In particular, using the thinned CRM framework, DRGPM can effectively activate newly-formed communities and switch off redundant communities over time, which strengthens the model interpretability for longitudinal network analysis. In Fig. 3.10, we depict the thinning probabilities (the mean of $b_k^{(t)}$) over time for the six inferred active communities by DRGPM. We notice that DRGPM infers three communities $(4, 11, 24)$ at $t = 0$, turns off Group 4 and turns on Groups 3 and 7 at $t = 10$. Group 28 is only active from $t = 30$ to 40. The inferred link probabilities and node-community associations $\phi_{uk}^{(t)}\Omega_{kk}$ by DPGM are depicted in columns (c) and (d) of Figure 3.8, respectively. We note that both DPGM and DRGPM infer fewer numbers of groups than D-GPPF because dynamic node-group connections are explicitly modelled by time-evolving node memberships in the former two methods. In particular, we notice that DPGM unavoidably generates some redundant groups that lack interpretability. This is due to that DPGM assumes the inferred group weights to be static throughout the whole time period.

3.4.3 *Experimental setup*

For the quantitative evaluation, we consider the following datasets:

1. **Face-to-face dynamic contacts network (FFDC):** This dataset (Mastrandrea et al. 2015) records timestamped face-to-face contacts among 180 students for 7 school days. We generated a dynamic network considering each school day as a snapshot, and created an edge

between each pair of students at time t if they have at least one contact recorded at that given time.

2. **DBLP:** The DBLP co-authorship network data (Asur et al. 2009) contains the co-authorship information among 958 authors over ten years (1997-2006) in 28 conferences which spans three related research topics-database, data mining, and artificial intelligence. We focus on a subset of 324 most connected authors over all time period.

3. **Enron:** The Enron data[4] contains 517,431 emails among 151 users over 38 months (from May 1999 to June 2002). We generated a dynamic network aggregating the data into monthly snapshots, and created an edge between each pair of users at time t if they have at least one email recorded at that given time.

The summary statistics are detailed in Table 3.1.

Dataset	FFDC	DBLP	Enron
# Nodes	180	324	151
# Slices	7	10	38
# Edges	8,332	11,154	11,414

Table 3.1: Details of the datasets used in our experiments.

3.4.4 *Predicting missing links*

First, we performed missing link prediction on the real-world datasets, and show the proposed model's predictive performance compared to the baseline models. We randomly held out 20% of the network entries (either links or non-links) for each snapshot as test data, and used the remaining 80% to predict the held-out entries. DRIFT was infeasible to run on the Enron dataset in a reasonable amount of time given our computational resource. For DSBM, we either set K to the true number of classes provided by the dataset or initialized it by examining the singular values of the first snapshot (K. S. Xu et al. 2014). We applied HGPEPM to each snapshot of dynamic networks independently. For all probabilistic methods, we ran 2000 burn-in iterations, and collected 1000 samples from the model posterior distribution. We estimated the posterior mean of the edge probability for each held-out edge in the test data by averaging over the collected Gibbs samples. We then used these edge probabilities to evaluate the predictive performance of each model by calculating the area under the curve of the receiver operating characteristic (AUROC) and of the precision-recall (PR). In Table 3.2, we report the average evaluation metrics for each model over 10 runs.

Overall, we found that DRIFT performs slightly better than DRGPM, although DRGPM has a significant advantage in terms of computational cost due to the Bernoulli-Poisson link. HGPEPM performs better than the dynamic models on the DBLP dataset because co-authorship links change dramatically from one year to the next one, and hence, the static model is better at fitting each snapshot independently. For the longitudinal Enron email network that is recorded monthly, DRGPM performs better than the baseline methods.

---

[4] `https://www.cs.cmu.edu/~enron/`.

	Missing Link Prediction		Future Network Forecasting	
Model	FFDC			
HGPEPM	0.917 ± 0.006	0.354 ± 0.018	0.733 ± 0.025	0.164 ± 0.022
DSBM	0.878 ± 0.011	0.251 ± 0.017	0.825 ± 0.085	0.181 ± 0.039
D-GPPF	0.908 ± 0.005	0.313 ± 0.019	0.842 ± 0.028	0.203 ± 0.046
DRIFT	**0.933** ± 0.006	**0.416** ± 0.020	0.848 ± 0.056	**0.224** ± 0.025
DPGM	0.921 ± 0.004	0.359 ± 0.020	0.846 ± 0.017	0.221 ± 0.036
DRGPM	0.924 ± 0.005	0.357 ± 0.018	**0.852** ± 0.033	**0.226** ± 0.040

	Missing Link Prediction		Future Network Forecasting	
Model	DBLP			
HGPEPM	**0.979** ± 0.004	**0.791** ± 0.014	0.714 ± 0.035	0.106 ± 0.027
DSBM	0.913 ± 0.006	0.256 ± 0.009	0.704 ± 0.030	0.091 ± 0.009
D-GPPF	0.914 ± 0.005	0.308 ± 0.018	0.734 ± 0.080	0.109 ± 0.046
DRIFT	0.970 ± 0.019	0.491 ± 0.025	0.745 ± 0.060	0.121 ± 0.054
DPGM	0.960 ± 0.002	0.423 ± 0.032	0.744 ± 0.053	**0.123** ± 0.064
DRGPM	0.963 ± 0.003	0.425 ± 0.023	**0.753** ± 0.057	**0.127** ± 0.053

	Missing Link Prediction		Future Network Forecasting	
Model	Enron			
HGPEPM	0.972 ± 0.001	0.443 ± 0.016	0.828 ± 0.073	0.246 ± 0.140
DSBM	0.916 ± 0.007	0.225 ± 0.023	0.853 ± 0.059	0.325 ± 0.116
D-GPPF	0.977 ± 0.002	0.499 ± 0.022	0.878 ± 0.057	**0.360** ± 0.121
DRIFT	NA	NA	NA	NA
DPGM	0.979 ± 0.002	0.565 ± 0.014	**0.883** ± 0.051	**0.361** ± 0.131
DRGPM	**0.983** ± 0.002	**0.597** ± 0.017	**0.886** ± 0.067	**0.363** ± 0.130

Table 3.2: Quantitive evalution. We highlight the performance of the best scoring model in bold.

3.4.5 *Forecasting future networks*

Next, we consider the task of forecasting an unseen network snapshot $A^{(t)}$ given observed snapshots $A^{(1:t-1)}$. Following previous works (J. R. Foulds et al. 2011; Heaukulani et al. 2013; M. Kim et al. 2013), we trained the models on the first $(t-1)$ snapshots of the considered network, and then estimated the predictive distribution of the unseen snapshot $A^{(t)}$ by running MCMC sampling one time step into the future. We applied HGPEPM to the most recent snapshot $A^{(t-1)}$, and then to perform prediction on the unseen snapshot $A^{(t)}$. For DRGPM, we set $\Omega^{(t)} = \Omega^{(t-1)}$, assuming the snapshots at nearby time points share a similar set of groups. We generated 10 samples of $Z^{(t)}$ for each of the 1000 samples collected for $Z^{(t-1)}$. For DSBM, we used the method detailed in (K. S. Xu et al. 2014) to perform future network forecasting. Table 3.2 shows the averaged performance for each model over different network snapshots from 3 to T. Overall, DRGPM shows competitive performance on all three datasets. This confirms that DRGPM can flexibly characterize temporally local links via time-evolving node memberships and switches off redundant groups to avoid overfitting the data.

3.4.6 *Running Time*

The probilistic models achieve higher accuracy although these methods require more computation time to collect MCMC samples. DSBM is much faster than the probabilistc models because its inference is performed using the extended Kalman filter. Table 5.2 compares the per-iteration computation time of the sampling-based models (all models are implemented in Matlab). The computational cost of DRIFT scales in $\mathcal{O}(\bar{K}^2V^2T)$, where $\bar{K}$ is the expected number of groups. The Bernoulli-Poisson link based models (D-GPPF, DPGM, DRGPM) are much faster than the logistic link based method (DRIFT) because the former models scale linearly with the number of non-zero entries in network data. The computational complexity of DRGPM, D-GPPF and DPGM are $\mathcal{O}(N^e\bar{K}^2 + V\bar{K}T + \bar{K}^2T)$, $\mathcal{O}(N^e\bar{K} + V\bar{K} + \bar{K}T)$ and $\mathcal{O}(N^e\bar{K}^2 + V\bar{K}T + \bar{K}^2)$, respectively. DRGPM is slightly faster than DPGM because DRGPM can effectively turn off redundant groups and hence achieves a lower computational cost.

	FFDC	DBLP	Enron
DRIFT	164.342	382.119	-
D-GPPF	0.145	0.242	0.292
DPGM	0.748	1.676	1.705
DRGPM	0.623	1.217	1.234

Table 3.3: Comparison of computation time (seconds per iteration).

3.4.7 *Case Study: Military Interstate Disputes Dataset*

We investigated the military interstate disputes (MID) dataset that contains disputes events between 138 countries from 1992 to 2001 (Ghosn et al. 2004) to explore the latent structure discovered by DRGPM. A dynamic network was generated by aggregating the data into monthly snapshots and a link was created between each pair of two countries if either country has disputes with the other one at that given time. We applied DRGPM to this dynamic network initializing $K = 30$ groups.

Most of the identified groups correspond to some regional relations or conflicts. In Figure 3.11, we depict four interesting groups inferred by DRGPM and show the group activity by plotting the mean of the thinning function $b_k^{(t)}$. We normalized the node memberships to $[0, 1]$ by dividing them by the sum of memberships within the same group. In Table 3.4, we report the top 20 nodes associated to each of four groups with positive memberships. For instance, we found that Group 1 corresponds to the second Congo war (1998-2000). The first six nodes of the group are indeed the belligerents of this war. Group 2 corresponds to the Bosnian War (1992-1995), and its associated nodes are Yugoslavia and some NATO members that are indeed the belligerents of this war. Groups 3 and 4 are related to the regional disputes between some African countries.

3.5 CONCLUSIONS

We proposed the dynamic Poisson gamma memberships model for discrete-time dynamic networks. The evolution of the underlying structure is characterized by the Markov construction of latent memberships. The new proposed framework can automatically infer an appropriate number of latent communities from network data. We also proposed efficient batch and online Gibbs algorithms that make use of the data augmentation and marginalization technique. To improve the model flexibility, we also generalized the DPGM to characterize the group birth/death dynamics using the thinned completely random measures (tCRMs), which enable us to investigate the evolution of the inferred latent structure. The inferred latent dynamic structure can be useful for various qualitative analysis in practical applications. We experimentally demonstrated the competitive predictive performance and scalability of our framework on three real-world datasets.

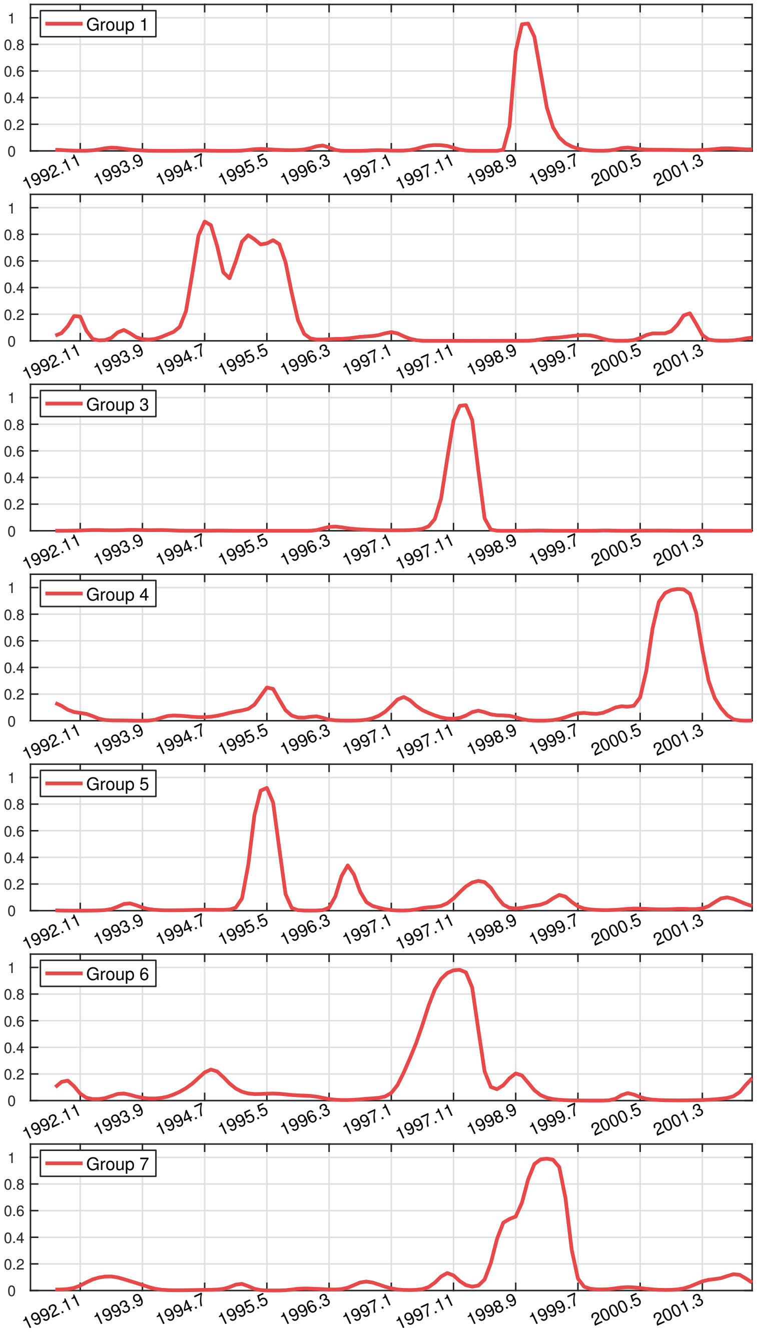


Figure 3.11: The activity (mean of $b_k^{(t)}$) of the selected groups $(1-7)$ inferred from the MID network.

Group	Country
1	Namibia (0.22), Chad (0.21), Zimbabwe (0.21), Angola (0.20), Dem. Rep. Congo (0.10), Sudan (0.05), Zambia (0.01)
2	Yugoslavia (0.60), Greece (0.13), Italy (0.04), UK (0.04), France (0.04), Belgium (0.03), Albania (0.03), Turkey (0.03), USA (0.02), Spain (0.02), Netherlands (0.01), Germany (0.01)
3	Nigeria (0.45), Ghana (0.31), Guinea (0.24)
4	Liberia (0.98), Sierra Leone (0.02)
5	Taiwan (0.47), China (0.20), Thailand (0.13), Philippines (0.10), Cambodia (0.05), Vietnam (0.03), Turkey (0.01) Togo (0.01)
6	Sierra Leone (0.90), Nigeria (0.04), Guinea (0.04), Ghana (0.02)
7	Norway (0.42), Canada (0.28), Portugal (0.14), Turkey (0.14), United States of America (0.02)
8	Uganda (0.57), Rwanda (0.41), Eritrea (0.01), Congo (0.01), Bahrain (0.01)
9	Yugoslavia (0.88), United States of America (0.03), Denmark (0.02), Russia (0.02), Canada (0.02), Haiti (0.01) Bangladesh (0.01), Cuba (0.01)
10	Iraq (0.45), North Korea (0.22), Russia (0.17), Cyprus (0.10), Greece (0.06)
11	United States of America (0.34), Turkey (0.25), United Kingdom (0.21), South Korea (0.12), Denmark (0.02) Trinidad and Tobago (0.02), Japan (0.02), Norway (0.01), Honduras (0.01)
12	Israel (0.98) , El Salvador (0.01) , United States of America (0.01)
13	Portugal (0.31), Turkey (0.17), United Kingdom (0.12), Denmark (0.11), Belgium (0.10), Norway (0.10), Albania (0.08)
14	South Korea (0.30), United States of America (0.22), Vietnam (0.18), Afghanistan (0.11), Norway (0.07), Mongolia (0.05), Denmark (0.04), Peru (0.03)
15	Turkey (0.47), United States of America (0.29), South Korea (0.09), Iran (0.08), Georgia (0.06)
16	Albania (0.25), Portugal (0.23), Canada (0.21), Denmark (0.18), Norway (0.13)

Table 3.4: The top 20 nodes associated with each of the selected groups as shown in Figs. 3.11 to 3.13 from the MID network. The highest node memberships to the corresponding selected groups throughout the whole period are reported for each node in parentheses.

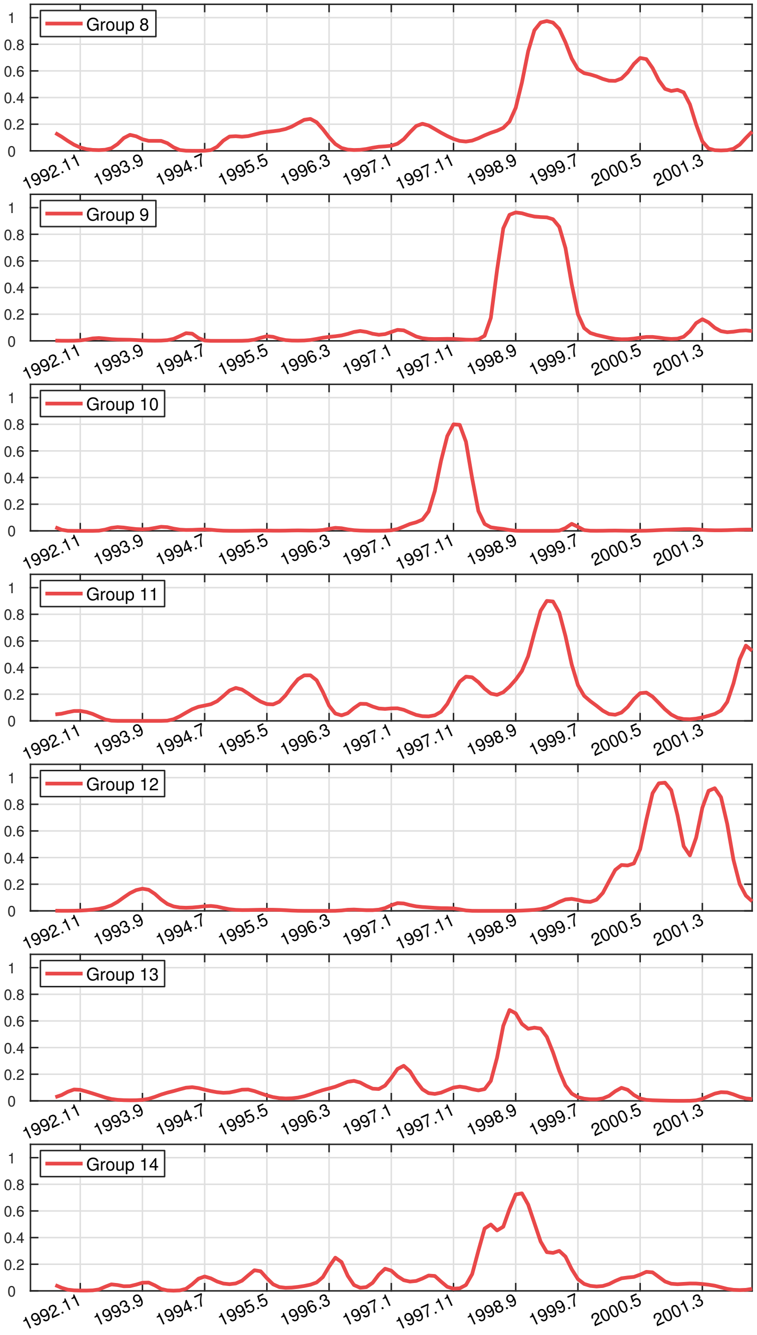


Figure 3.12: The activity (mean of $b_k^{(t)}$) of the selected groups $(8-14)$ inferred from the MID network.

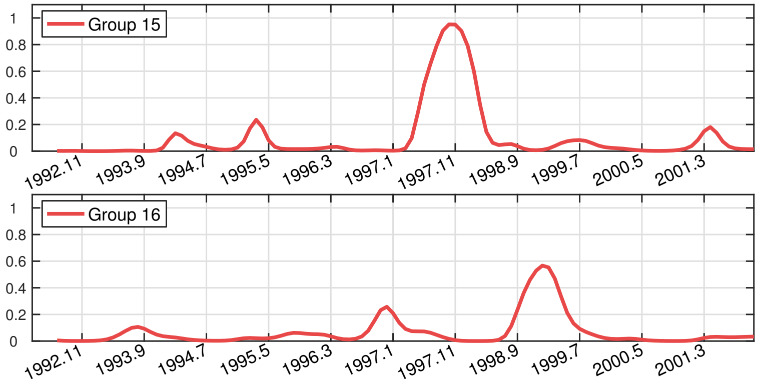


Figure 3.13: The activity (mean of $b_k^{(t)}$) of the selected groups $(15-16)$ inferred from the MID network.

4

CONTINUOUS TIME DYNAMIC NETWORK MODELS

There is an increasing interest in modeling and understanding the information diffusion pathways and interaction dynamics among entities from continuously generated streams of data. These streaming data include the timestamped interaction events among entities (e.g., question-answering threads (Mavroforakis, Valera, and Gomez-Rodriguez 2017), email communications (J. Yang et al. 2017) and interaction events among nations (Schein et al. 2016a)), and the auxiliary contents created by these interacting entities. Such temporal interaction data enable us not only to track the *topics* underlying the human-generated contents, but also to understand the network formation and evolving process among these interacting entities.

The Hawkes process-based models (Blundell, Beck, and Heller 2012; Du et al. 2015b; H. Xu et al. 2016a; J. Yang et al. 2017; Miscouridou et al. 2018) received a lot of attention as Hawkes processes (Hawkes 1971) are particularly well fitted to model the inherent self-excitatory (a single point process) or mutual-excitatory nature revealed by many real-world temporal interactions. Recent work (Blundell, Beck, and Heller 2012; DuBois et al. 2013; Linderman and Adams 2014; Junuthula et al. 2017; J. Yang et al. 2017; Miscouridou et al. 2018) attempt to hybridize statistical models for *static* networks with Hawkes processes to model both *implicit* social structure and *reciprocity* among entities from their temporal interactions. Among these methods, the Hawkes infinite relational model (Hawkes-IRM) (Blundell, Beck, and Heller 2012) characterizes the interaction dynamics between groups of individuals using mutually-exciting Hawkes processes. To further capture the reciprocating interactions between individuals, (Miscouridou et al. 2018) proposes to incorporate the underlying overlapping community structure into the base intensity of the Hawkes processes, via the compound completely random measure (CCRM) prior (Todeschini et al. 2017). Despite having many attractive properties, the Hawkes-CCRM is restrictive in that the reciprocity in all interactions are captured via the same triggering kernel, and thus fails to interpret the differences in interaction dynamics across individuals. For example, an employee may reply back to the emails from his/her department more quickly than responding to non-urgent emails from outside. A fundamental problem in modeling such temporal dynamics is to infer the *latent struture* behind observed events (Du et al. 2015a; Mavroforakis, Valera, and Gomez-Rodriguez 2017; H. Xu and Zha 2017; Tan, Rao, and Neville 2018).

In this chapter, we attempt to develop a new framework, the Hawkes edge partition model (Hawkes-EPM) , which hybridizes the recently advanced hierarchical gamma process edge partition model (HGP-EPM) (M. Zhou et al. 2015) with Hawkes processes. More specifically, the base intensity of the Hawkes process is built upon the latent representations inferred by the HGP-EPM, which enables us to capture the overlapping communities, degree heterogeneity and sparsity underlying the observed interactions. To accurately capture the interaction dynamics between two individuals, our model augments each specific interaction between them with a pair of latent variables, to indicate which of their latent communities (features) leads to the occurring of that interaction. Accordingly, the excitation effect of each interaction on its opposite direction is determined by its latent variables. For instance, as shown in Figure 4.1, Bob and Helen have many common interests (features), and some of their interactions are due to their common interests in playing baseball. Moreover, our model estimates the number of the underlying communities via the inherent shrink-

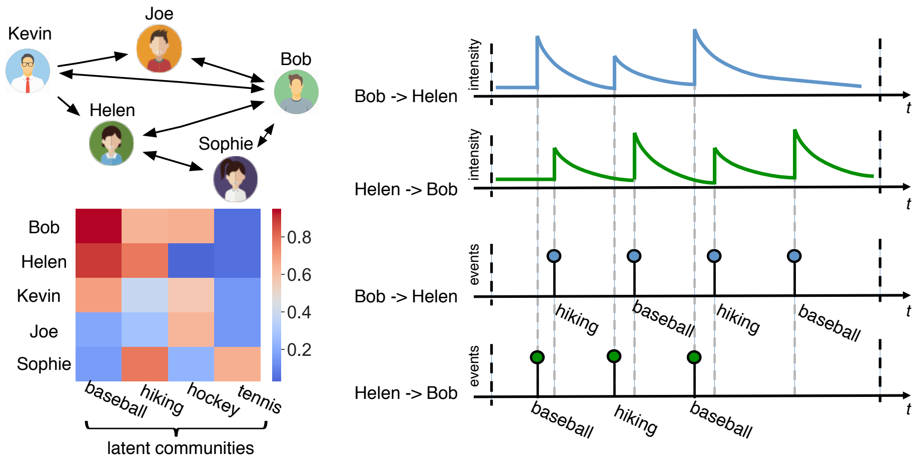


Figure 4.1: An illustrative example. The top left figure plots the aggregated *directed* networks from the temporal interactions among five nodes. The bottom left graph shows the underlying community structure. We see that both Bob and Helen have interest in "baseball" and "hiking". The top right graph plots the intensity functions of the interactions from Bob to Helen, and from Helen to Bob, respectively. The bottom right graph plots the interaction events from Bob to Helen, and from Helen to Bob, respectively. These interactions may represent the messages communicated between the involved nodes. As in this example, some of their interactions are about "baseball", and others relate to "hiking". We assume that behind each interaction, the latent patterns of the involved nodes determines the excitation effects of that event on the opposite direction.

age mechanism of the hierarchical gamma process (M. Zhou and L. Carin 2015a). Furthermore, our model construction can flexibly incorporate the auxiliary individuals' attributes, or covariates associated with interaction events.

The rest of this chapter is organized as follows. Section 4.1 discusses how the proposed model relates to previous work. Section 4.2 presents the Hawkes-edge partition model (Hawkes-EPM) model for continuous-time dynamic networks. In Section 4.3, we develop simple and efficient Gibbs sampling and Expectation-Maximization inference algorithms for the proposed Hawkes-EPM model. Section 4.4 presents the experimental results of the two developed models compared with state-of-the-art methods on both synthetic and real-world datasets. Finally, Section 4.5 concludes the chapter.

4.1 RELATED WORK

The proposed model closely relates to the Hawkes process-based interaction models and the Bayesian nonparametric prior-based Hawkes process models.

Hawkes processes-based Interaction Models. (Blundell, Beck, and Heller 2012) developed the Hawkes infinite relational model (Hawkes-IRM), in which the interaction dynamics between two nodes are determined by their respective community intensities while Hawkes-IRM allows each node to be affiliated with only one community (non-overlapping). Along this line of research, (Tan, Rao, and Neville 2018) proposed an Indian Buffet Hawkes process model, which assumes that each event can be simultaneously driven by multiple evolving factors shared by past events. (Miscouridou et al. 2018) developed the Hawkes-compound completely random measure model

(Hawkes-CCRM) by modelling the base rate of the Hawkes process via the CCRM model, which allows the overlapping community structure, sparsity and degree heterogeneity to be captured. Here, the proposed model not only models the interpretable latent structure underlying observed interactions as in (Miscouridou et al. 2018), but also captures the latent pattern behind each event.

Bayesian Nonparametric Hawkes Processes (BNHPs). Recently, Bayesian nonparametric priors (BNPs) (Ferguson 1973) are introduced to capture the latent structure underlying the observed event sequence. The Dirichlet-Hawkes process (DHP) (Du et al. 2015a) models the latent clustering structure underlying the observed events using the Dirichlet process. The Indian buffet Hawkes process (Tan, Rao, and Neville 2018) and the nested Chinese Restaurant process-Hawkes process (NCRP-HP) (Tan et al. 2018) have been developed to capture the rich factor-structured and hierarchically-structured temporal dynamics, respectively. (Mavroforakis, Valera, and Gomez-Rodriguez 2017) points out that most previous BNHP models suffer from the vanishing prior problem as the instantiated patterns in these models are only captured via the endogenous intensity. Hence, an already used pattern will vanish if its intensity tend to be zero. As a consequence, these BNHP methods unavoidably generate many redundant patterns for the events widely separated in time but share similar dynamics. (Mavroforakis, Valera, and Gomez-Rodriguez 2017) resolved this issue using the hierarchical Dirichlet process (Teh et al. 2007) framework, where the top-layer Dirichlet process defines the distribution over latent patterns, and the bottom-layer Hawkes processes capture the temporal dynamics across multiple event sequences. Our proposed model infers the appropriate number of communities (patterns) using the hierarchical gamma process prior (M. Zhou and L. Carin 2015a). In the Hawkes-EPM, each latent pattern is modelled by a community-specific intensity function, which is non-negligible over time, and thus effectively prevents from the vanishing prior issue.

4.2 HAWKES PROCESSES WITH THE HIERARCHICAL GAMMA PROCESS EDGE PARTITION MODEL

Next we shall briefly review the hierarchical gamma process edge partition model (M. Zhou et al. 2015), and then introduce the Hawkes-EPM for temporal interaction events.

4.2.1 *Hierarchical Gamma Process Edge Partition Models*

The hierarchical gamma process edge partition (HGP-EPM) model (M. Zhou et al. 2015) was recently proposed to detect overlapping community structure in static relational data. Formally, let $\mathcal{V}$ denotes a set of nodes, and the (static) relationships among $V \equiv |\mathcal{V}|$ nodes be represented by a binary adjacency matrix $E \in \{0,1\}^{V \times V}$, where $e_{uv} = 1$ if there is an (directed) edge from nodes u to v, and 0 otherwise. We ignore self-edges $\{e_{uu}\}_{u \in \mathcal{V}}$ as a node never interacts with itself. The (truncated) HGP-EPM is generated as

$$\phi_{u,k} \sim \text{Gamma}(a_u, \frac{1}{c_u}), \quad a_u \sim \text{Gamma}(e_0, \frac{1}{f_0}), \quad r_k \sim \text{Gamma}(\frac{r_0}{K}, \frac{1}{c_0}), \tag{4.2.1}$$

$$\Omega_{k,k'} \sim \begin{cases} \text{Gamma}(\xi r_k, \chi), & \text{if } k = k' \\ \text{Gamma}(r_k r_{k'}, \chi), & \text{otherwise} \end{cases}, \; e_{u,v} \sim \text{Bernoulli}\left[1 - \prod_{k,k'=1}^{K} \exp(-\phi_{u,k}\Omega_{k,k'}\phi_{v,k'})\right],$$

where each node $u \in \mathcal{V}$ is chacterized by a *positive* feature vector $(\phi_{u,1}, \ldots, \phi_{u,K})$ with $\phi_{u,k}$ measuring how strongly node u is affiliated to each community $k = 1, \ldots, K$. a_u captures the *sociabil-*

ity of node u, and thus node u exhibiting a large number of interactions will be characterized by a large a_u. The prevalence of each community k is captured by a positive weight r_k. The HGP-EPM can infer an appropriate number of communities via its inherent shrinkage mechanism: many communities' weights $\{r_k\}$ tend to be small as $K \to \infty$, and thus most redundant communities will be shrunk effectively. The parameters $\Omega_{k,k}, \Omega_{k,k'}$ captures the intra-community and inter-community interaction weights, respectively. The probability of there being an edge from node u to node v is parameterized under the Bernoulli-Poisson link (BPL) function $\Pr(y = 1 \mid \zeta) = 1 - e^{-\zeta}$, where ζ defines the positive rate. We note that the HGP-EPM using the BPL function well fits large *sparse* graphs (M. Zhou 2018a). Following (M. Zhou et al. 2015), we impose the $\mathrm{Gamma}(1, 1)$ prior over the hyperparameters $c_u, c_0, e_0, f_0, r_0, \xi, \chi$. Interestingly, the probability of an edge e_{uv} modeled by the BPL can be equivalently generated as

$$e_{u,v} = \mathbb{1}(\tilde{e}_{u,v} \geq 1), \quad \tilde{e}_{u,v} \sim \text{Poisson}\left(\sum_{k=1}^{K} \sum_{k'=1}^{K} \phi_{u,k} \Omega_{k,k'} \phi_{v,k'}\right),$$

where $\phi_{u,k} \Omega_{k,k'} \phi_{v,k'}$ captures the connecting strength between nodes u and v due to their affiliations to communities k, k', respectively. Note that the HGP-EPM not only captures the *overlapping* community structure, degree *heterogeneity*, but also characterizes both *homophily* and *stochastic equivalence* exhibited in real-world interactions (M. Zhou 2018a).

4.3 THE HAWKES EDGE PARTITION MODEL

Let $\{(t_i, s_i, d_i)\}_{i=1}^{N}$ be a sequence of temporal interaction events, where (t_i, s_i, d_i) denotes a *directed* interaction from node s_i (sender) to node d_i (receiver) at time t_i. Following (Miscouridou et al. 2018), we build the non-time dependent component μ_{s_i, d_i} of the intensity function upon the latent parameters inferred by the HGP-EPM. More specifically, for the reciprocating Hawkes process from node u to node v, we let $\mu_{u,v} = \sum_{k,k'} \phi_{u,k} \Omega_{k,k'} \phi_{v,k'}$, where $\phi_{u,k}$ captures the affiliation of node u to community k, and $\Omega_{k,k'}$ the inter-community interaction strength between k and k'. Hence, our base rate naturally models that two nodes sharing more features are more likely to interact with each other. Different from the Hawkes-CCRM, we further assume that the interaction dynamics between nodes s_i and d_i are also influenced by their respective affiliated communities. To do so, for each interaction event, we introduce the latent variables z_i^s and z_i^d to represent the latent patterns of the sender s_i and the receiver d_i in i-th event, respectively. Inferring the latent patterns underlying each event is the key to accurately characterize the interaction dynamics between entities involved. More specifically, we define the intensity function for a pair of nodes u and v as

$$\begin{aligned} \lambda_{u,v}(t) &= \mu_{u,v} + \sum_{j: t_j \in \mathcal{H}_{v,u}(t)} \gamma_{z_j^d, z_j^s}(t - t_j) \\ &= \sum_{k,k'} \left\{ \mu_{u,k,k',v} + \sum_{j: t_j \in \mathcal{H}_{v,u}(t) \mid z_j^s = k', z_j^d = k} \alpha_{kk'} \exp[-(t - t_j)/\delta] \right\}, \end{aligned} \tag{4.3.2}$$

where we define the base rate $\mu_{u,v} \equiv \sum_{k,k'} \mu_{u,k,k',v} \equiv \sum_{k,k'} \phi_{u,k} \Omega_{k,k'} \phi_{v,k'}$. Here $\mu_{u,k,k',v}$ accounts for the exogenous interactions from u to v due to their respective affiliations to k, k', respectively. $\gamma_{kk'}(t)$ is a nonnegative kernel function that captures the decaying influence of past events under the pattern (k', k) on the current intensity. In this chapter, we assume that the current rate

$\lambda_{u,k,k',v}(t)$ from u to v under the pattern (k, k') is only influenced by the past opposite interactions $\{(t_j, s_j, d_j) \mid t_j < t, s_j = v, d_j = u\}$ under the pattern (k', k), which we denote by $\{t_j \in \mathcal{H}_{v,u}(t) \mid z_j^s = k', z_j^d = k\}$. More specifically, $\alpha_{kk'}$ controls the excitatory effect under the pattern (k', k), and we impose a gamma prior over $\alpha_{kk'}$, i.e., $\alpha_{kk'} \sim \mathrm{Gamma}(e_0, 1/f_0)$. Figure 4.2 presents a simple illustrative example for the Hawkes-EPM.

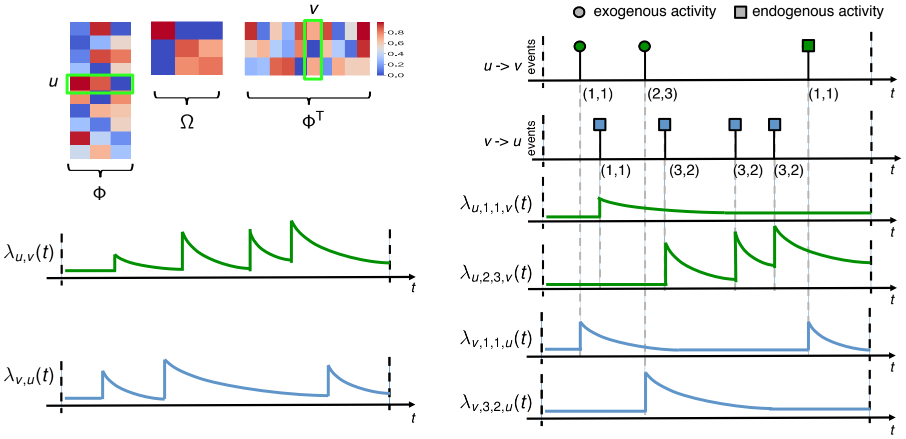


Figure 4.2: A simple example for the Hawkes-EPM model. The top left figure shows the inferred matrix of node features Φ, and the community-community interaction strength Ω. Here, node u connects to node v through the intra-community interaction $(1, 1)$ and inter-community interaction $(2, 3)$. The top right figure plots the interaction events between u and v. Each event is denoted by a bar, under which we use (a, b) to indicate the latent variables a, b of nodes u, v in that event, e.g., $z_1^u = 1, z_1^v = 1$ for 1-st event. The bottom left figure plots the intensities of the interactions from u to v, and from v to u, respectively. Equivalently, $\lambda_{u,v}(t)$ can be represented by the summation of $\{\lambda_{u,k,k',v}(t)\}_{k,k'}$, where $\lambda_{u,k,k',v}(t)$ denotes the interaction intensity from u to v via the inter-community (k, k').

If (t_i, s_i, d_i) is an *exogenous* event induced by μ_{s_i,d_i}, the latent patterns (z_i^s, z_i^d) for (s_i, d_i) are determined by their affiliated communities via $(\boldsymbol{\phi}_{s_i}, \boldsymbol{\phi}_{d_i})$, respectively. In case that (t_i, s_i, d_i) is an *endogenous* event, (z_i^s, z_i^d) are determined by the past opposite interactions from d_i to s_i. More specifically, the latent patterns associated to i-th event can be generated as

$$\begin{aligned}
&\Pr(z_i^s = k, z_i^d = k' \mid t_i, s_i = u, d_i = v) \\
&= \left(\mu_{u,k,k',v} + \sum_{j: t_j \in \mathcal{H}_{v,u}(t) \mid z_j^s = k', z_j^d = k} \alpha_{kk'} \exp[-(t_i - t_j)/\delta] \right) / \lambda_{u,v}(t_i), \quad \text{for } k, k' \in 1,\ldots,K.
\end{aligned} \tag{4.3.3}$$

In real temporal interactions, some additional information such as auxiliary node attributes, explicitly declared relationships among entities, and communicating contents are also available for accurately modelling temporal interaction dynamics when interaction events are incomplete (say, due to the privacy issues of individuals). Formally, let $\mathbf{x}_{u,v} \equiv (x^1_{u,v}, \dots, x^D_{u,v})^{\mathrm{T}}$ denotes the covariates of D dimensions associated with a pair of nodes u and v. For example, the covariates $\mathbf{x}_{u,v}$ may represent the common attributes shared by u and v, or the word embeddings inferred from the communicating contents between u and v. We generalize the Hawkes-EPM model by letting

$$\mu_{u,k,k',v} \sim \mathrm{Gamma}(\tilde{\mu}_{u,k,k',v}, 1/(\exp[-\mathbf{x}^{\mathrm{T}}_{u,v}\boldsymbol{\beta}_{kk'}])), \tag{4.3.4}$$

where $\tilde{\mu}_{u,k,k',v} \equiv \phi_{u,k}\Omega_{k,k'}\phi_{v,k'}$, and $\boldsymbol{\beta}_{kk'} \equiv (\beta^1_{k,k'}, \dots, \beta^D_{k,k'})^{\mathrm{T}}$ is the regression coefficient vector of latent pattern (k, k'). The base intensity in (4.3.4) is drawn from a gamma prior where the shape parameter incorporates the underlying community structure information via $\tilde{\mu}_{u,k,k',v}$, and the scale parameter is a function of the input auxiliary covariates. To our knowledge, the regression component in (4.3.4) is investigated already in (M. Zhou 2018b; Rai et al. 2015; Q. Zhang and M. Zhou 2018), but firstly applied in this context.

Remarks. Note that the proposed model allows an unbounded number of latent patterns to be shared across all pairs of interacting nodes via the hierarchical gamma process (HGP) (M. Zhou and L. Carin 2015a). As shown in (4.3.3), the base rate $\mu_{u,k,k',v}$ of the latent pattern (k, k') is non-negligible over the whole time period, and thus our model allows the events widely separated in time but with similar dynamics to be parameterized under the same pattern, to avoid *vanishing* prior issue (Mavroforakis, Valera, and Gomez-Rodriguez 2017; Kapoor et al. 2018).

4.4 INFERENCE

The proposed model admits efficient approximate inference as the posteriors of all the model parameters are available in closed-form using Pólya-Gamma data augmentation strategy. Let $\mathcal{D}$ denotes the whole events data, E the *binary* adjacency matrix aggregated from $\mathcal{D}$, i.e., $e_{uv} = 1$ for $u, v \in \mathcal{V}$ if there being at least one interaction observed in the time interval $[0, T]$, Ξ the parameters of the HGP-EPM, and Θ the parameters of the Hawkes-EPM. We use the "$\hat{x}$" to denote the maximum a posterior (MAP) estimate of x. Following (Miscouridou et al. 2018), we present a two-step inference procedure: (i) Approximate $\Pr(\Xi \mid \mathcal{D}, E)$ by $\Pr(\Xi \mid E)$, and obtain a maximum a posterior estimate $\hat{\Xi}$, and then (ii) Approximate $\Pr(\Theta \mid \Xi, \mathcal{D})$ by $\Pr(\Theta \mid \hat{\Xi}, \mathcal{D})$. The full posterior is approximated by $\Pr(\Theta, \Xi) = \Pr(\Xi \mid E)\Pr(\Theta \mid \hat{\Xi}, \mathcal{D})$. The posterior inference for $\hat{\Xi}$ is performed using the Gibbs sampling procedure described in (M. Zhou et al. 2015). The inference algorithm for the hierarchical gamma process edge partition model (HGP-EPM) is detailed in (M. Zhou et al. 2015) with the released code[1]. Next we shall explain the Gibbs sampling and Expectation-Maximization algorithms to infer the parameters of the Hawkes-EPM.

4.4.1 *Gibbs Sampling*

Sampling latent variables $\{z^s_i, z^d_i\}_{i=1}^N$: For each event (t_i, s_i, d_i), we utilize an auxiliary binary variable b_i to denote whether i-th event is triggered by the base rate (exogenous) or by opposite past interactions (endogenous) as

$$(b_i \mid -) \sim \mathrm{Bernoulli}(\mu_{s_i,d_i} / \lambda_{s_i,d_i}(t_i)). \tag{4.4.1}$$

---

[1] `https://github.com/mingyuanzhou/EPM`.

Then, we sample the latent patterns (z_i^s, z_i^d) for each event as

$$(z_i^s, z_i^d \mid -) \sim \begin{cases} \text{Categorical}\Big(\frac{\{\mu_{s_i,k,k',d_i}\}_{k,k'=1}^K}{\lambda_{s_i,d_i}(t_i)}\Big), & \text{if } b_i = 1 \\ \text{Categorical}\Big(\frac{\{\lambda^{\text{endo}}_{s_i,k,k',d_i}(t_i)\}_{k,k'=1}^K}{\lambda_{s_i,d_i}(t_i)}\Big), & \text{otherwise} \end{cases} \tag{4.4.2}$$

where we define

$$\lambda^{\text{endo}}_{s_i,k,k',d_i}(t_i) \equiv \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t) | z_j^s = k', z_j^d = k} \alpha_{kk'} \exp[-\delta(t_i - t_j)]. \tag{4.4.3}$$

Given the sampled latent variables, we update the sufficient statistics as

$$\begin{aligned} m^{\text{exo}}_{u,k,k',v} &\equiv \sum_j \mathbf{1}(b_j = 1, s_j = u, d_j = v, z_j^s = k, z_j^d = k'), \\ m^{\text{endo}}_{u,k,k',v} &\equiv \sum_j \mathbf{1}(b_j = 0, s_j = u, d_j = v, z_j^s = k, z_j^d = k'). \end{aligned} \tag{4.4.4}$$

The log-posterior of the observed temporal events $\mathcal{D} \equiv \{(t_i, s_i, d_i)\}_{i=1}^N$ is

$$\begin{aligned} \mathcal{L}(\Theta) = &\textstyle\sum_i \log \left\{ \mu_{s_i,d_i} + \sum_{k,k'} \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i) | z_j^s = k', z_j^d = k} \alpha_{kk'} \exp\left[-(t_i - t_j)/\delta\right] \right\} \\ &-\textstyle\sum_i \left\{ \mu_{s_i,d_i} T + \sum_{k,k'} \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i) | z_j^s = k', z_j^d = k} \alpha_{kk'} \delta(1 - \exp\left[-(t_i - t_j)/\delta\right]) \right\} + \log \Pr(\Theta). \end{aligned} \tag{4.4.5}$$

Sampling the kernel parameters $\{\alpha_{kk'}\}$: As we place gamma priors over $\alpha_{kk'}$ as $\alpha_{kk'} \sim \text{Gamma}(e_0, 1/f_0)$, and thus we have

$$(\alpha_{kk'} \mid -) \sim \text{Gamma}\left(e_0 + m^{\text{endo}}_{\cdot k,k' \cdot}, 1/\left(f_0 + \sum_i \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} \frac{1}{\delta} \left(1 - \exp\left[-\frac{(T - t_j)}{\delta} \right] \right) \right) \right), \tag{4.4.6}$$

where $m^{\text{endo}}_{\cdot k,k' \cdot} \equiv \sum_i m^{\text{endo}}_{s_i,k,k',d_i}$, and $m^{\text{endo}}_{\cdot k,k' \cdot}$ denotes the total number of endogenous events associated with the latent pattern (k, k').

Sampling the base intensity $\{\mu_{u,k,k',v}\}$: As we have gamma prior over $\mu_{u,k,k',v}$ as $\mu_{u,k,k',v} \sim \text{Gamma}(\tilde{\mu}_{u,k,k',v}, 1/(\exp[-\mathbf{x}_{u,v}^{\mathrm{T}} \boldsymbol{\beta}_{kk'}]))$, we have

$$(\mu_{u,k,k',v} \mid -) \sim \text{Gamma}\left(\tilde{\mu}_{u,k,k',v} + m^{\text{exo}}_{u,k,k',v}, 1/(T + \exp[-\mathbf{x}_{u,v}^{\mathrm{T}} \boldsymbol{\beta}_{kk'}]) \right), \tag{4.4.7}$$

Marginalizing out $\mu_{u,k,k',v}$ from the likelihood leads to

$$\begin{aligned} \Pr(\mathcal{D} \mid \mathbf{x}_{u,v}, \boldsymbol{\beta}_{kk'}) &= \int \Pr(\mathcal{D} \mid \mu_{u,k,k',v}) \Pr(\mu_{u,k,k',v} \mid \mathbf{x}_{u,v}, \boldsymbol{\beta}_{kk'}) \mathrm{d}\mu_{u,k,k',v} \\ &\propto \text{NB}(m^{\text{exo}}_{u,k,k',v}; \tilde{\mu}_{u,k,k',v}, \sigma[\mathbf{x}_{u,v}^{\mathrm{T}} \boldsymbol{\beta}_{kk'} + \log(T)]), \end{aligned}$$

where $\sigma(x) = 1/(1+\exp(-x))$ denotes the logistic function. Using the Pólya-Gamma data augmentation strategy (M. Zhou et al. 2012; Polson et al. 2013), we first sample

$$\begin{aligned}
(\omega_{u,k,k',v} \mid -) &\sim \mathrm{PG}(\mu_{u,k,k',v} + m^{\mathrm{exo}}_{u,k,k',v}, \psi_{u,k,k',v}), \\
(\psi_{u,k,k',v} \mid -) &\sim \mathcal{N}(\mu_{\psi}, \sigma_{\psi}),
\end{aligned} \tag{4.4.8}$$

where PG denotes a Pólya-Gamma draw, and where

$$\begin{aligned}
\psi_{u,k,k',v} &\equiv \mathbf{x}_{uv}^{\mathrm{T}} \boldsymbol{\beta}_{kk'} + \log(T\pi_{uv}), \\
\pi_{uv} &\sim \log\mathcal{N}(0, \tau^{-1}), \\
\sigma_{\psi} &= [\omega_{u,k,k',v} + \tau]^{-1}, \\
\mu_{\psi} &= \sigma_{\psi}\left[(m^{\mathrm{exo}}_{u,k,k',v} - \mu_{u,k,k',v})/2 + \tau(\mathbf{x}_{uv}^{\mathrm{T}} \boldsymbol{\beta}_{kk'} + \log(T))\right],
\end{aligned} \tag{4.4.9}$$

where $\log\mathcal{N}(\cdot)$ denotes the lognormal distribution.
Sampling the regression coefficients $\{\boldsymbol{\beta}_{kk'}\}$: Given $\{\boldsymbol{\psi}_{kk'} \equiv (\psi_{1kk'1}, \ldots, \psi_{Ukk'V})\}$, we sample $\{\boldsymbol{\beta}_{kk'}\}$ as

$$(\boldsymbol{\beta}_{k,k'} \mid -) \sim \mathcal{N}(\boldsymbol{\mu}_{\beta}, \boldsymbol{\Sigma}_{\beta}), \tag{4.4.10}$$

where $\boldsymbol{\Sigma}_{\beta} = (\tau \mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{A})^{-1}$, $\mathbf{A} \equiv \mathrm{diag}[\nu_1^{-1}, \ldots, \nu_D^{-1}]$, $\boldsymbol{\mu}_{\beta} = \tau \boldsymbol{\Sigma}_{\beta} \mathbf{X}^{\mathrm{T}} (\boldsymbol{\psi}_{kk'} - \log(T))$, and $\mathbf{X} \equiv [\mathbf{x}_{11}, \ldots, \mathbf{x}_{UV}]^{\mathrm{T}}$.

The full procedure of our Gibbs sampler is summarized in Algorithm 3.

4.4.2 *Expectation-Maximization*

To scale up the inference procedure for the proposed model, we also develop an efficient Expectation-Maximization (EM) algorithm to perform maximum a posterior (MAP) estimation following (Lewis and Mohler 2011; K. Zhou, Zha, and Song 2013; H. Xu et al. 2016b). More specifically, let $\Theta^{(l)}$ denotes the current model parameters, we construct a tight upper-bound of log-posterior in (4.4.5) via the Jensen's inequality as

$$\begin{aligned}
\mathcal{Q}(\Theta \mid \Theta^{(l)}) = &-\textstyle\sum_i \left\{\mu_{s_i,d_i} T + \sum_{k,k'} \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} \gamma_{kk'}(t_i - t_j)\right\} \\
&+ \textstyle\sum_i \sum_{k,k'} p^{\mathrm{exo}}_{ikk'} \log\left[\dfrac{\mu_{s_i,k,k',d_i}}{p^{\mathrm{exo}}_{ikk'}}\right] + \sum_{k,k'} \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} p^{\mathrm{endo}}_{ikk'} \left[\dfrac{\gamma_{kk'}(t_i - t_j)}{p^{\mathrm{endo}}_{ikk'}}\right] + \log \Pr(\Theta),
\end{aligned}$$

where

$$\begin{aligned}
p^{\mathrm{exo}}_{ikk'} &= \frac{\mu^{(l)}_{s_i,k,k',d_i}}{\left[\mu^{(l)}_{s_i,d_i} + \sum_{k,k'} \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} \gamma^{(l)}_{kk'}(t_i - t_j)\right]}, \\
p^{\mathrm{endo}}_{ikk'} &= \frac{\left[\sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} \gamma^{(l)}_{kk'}(t_i - t_j)\right]}{\left[\mu^{(l)}_{s_i,d_i} + \sum_{k,k'} \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} \gamma^{(l)}_{kk'}(t_i - t_j)\right]}.
\end{aligned} \tag{4.4.11}$$

The introduced variable $p^{\mathrm{exo}}_{ikk'}$ can be interpreted as the probability that i-th event is drawn from the base rate under the latent pattern (k, k'). $p^{\mathrm{endo}}_{ikk'}$ is the probability that i-th event is triggered by the opposite interaction events under the pattern (k', k). Accordingly, we update the sufficient statistics as

$$m^{\text{exo}}_{u,k,k',v} \equiv \sum_{i:s_i=u,d_i=v} p^{\text{exo}}_{ikk'}, \tag{4.4.12}$$

$$m^{\text{endo}}_{u,k,k',v} \equiv \sum_{i:s_i=u,d_i=v} p^{\text{endo}}_{ikk'}. \tag{4.4.13}$$

Expectations of Pólya-Gamma random variables are available in closed-form (Scott and Sun 2013), and given by

$$\mathsf{E}\left[\omega^{(l+1)}_{u,k,k',v}\right] = \left(\frac{\tilde{\mu}^{(l)}_{u,k,k',v} + m^{\text{exo}}_{u,k,k',v}}{2\psi^{(l)}_{u,k,k',v}}\right) \tanh\left(\frac{\psi^{(l)}_{u,k,k',v}}{2}\right). \tag{4.4.14}$$

Maximizing $\mathcal{Q}(\Theta)$ with respect to each of the model parameters $\{\mu_{u,k,k',v}\}$, $\{\alpha_{k,k'}\}$, $\{\boldsymbol{\beta}_{k,k'}\}$, $\{\boldsymbol{\psi}_{k,k'}\}$ fixing the rest, leads to closed-form updates for each of these. We update the remaining parameters as follows

$$\mu^{(l+1)}_{u,k,k',v} = \frac{(\tilde{\mu}_{u,k,k',v} + m^{\text{exo}}_{u,k,k',v})}{(T + \exp[-\mathbf{x}^{\mathrm{T}}_{u,v}\boldsymbol{\beta}^{(l)}_{kk'}])}, \tag{4.4.15}$$

$$\alpha^{(l+1)}_{kk'} = \frac{(e_0 + \sum_{u,v} m^{\text{endo}}_{u,k,k',v})}{\left(f_0 + \sum_i \sum_{j:t_j \in \mathcal{H}_{d_i,s_i}(t_i)} \delta\left(1 - \exp\left[-\frac{(T-t_j)}{\delta}\right]\right)\right)}, \tag{4.4.16}$$

$$\boldsymbol{\psi}^{(l+1)}_{k,k'} = \left[\text{diag}(\mathsf{E}\left[\boldsymbol{\omega}^{(l)}_{k,k'}\right]) + \tau\mathbf{I}\right]^{-1} \left[\frac{\tilde{m}^{\text{exo}}_{k,k'} - \boldsymbol{\mu}^{(l)}_{k,k'}}{2} + \tau(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^{(l)}_{k,k'} + \log(T))\right], \tag{4.4.17}$$

$$\boldsymbol{\beta}^{(l+1)}_{k,k'} = \tau(\tau\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^{\mathrm{T}}\left(\boldsymbol{\psi}^{(l)}_{kk'} - \log(T)\right), \tag{4.4.18}$$

where $\boldsymbol{\omega}^{(l)}_{k,k'} \equiv (\omega^{(l)}_{1,k,k',1}, \ldots, \omega^{(l)}_{U,k,k',V})$, $\tilde{m}^{\text{exo}}_{k,k'} \equiv (m^{\text{exo}}_{1,k,k',1}, \ldots, m^{\text{exo}}_{U,k,k',V})^{\mathrm{T}}$, and $\mathbf{A} \equiv \text{diag}[\nu_1^{-1}, \ldots, \nu_D^{-1}]$, $\boldsymbol{\mu}^{(l)}_{k,k'} \equiv (\mu^{(l)}_{1,k,k',1}, \ldots, \mu^{(l)}_{U,k,k',V})^{\mathrm{T}}$, and $\mathbf{X} \equiv [\mathbf{x}_{11}, \ldots, \mathbf{x}_{UV}]^{\mathrm{T}}$.

The full procedure of our EM algorithm is summarized in Algorithm 4.

Computational Cost. For the second inference step, computing the latent variables $\{z^s_i, z^d_i\}$ and updating the intensities for all the given events takes $\mathcal{O}(NK^2)$ time, where K is the estimated number of communities by HGP-EPM. Estimating $\{\alpha_{kk'}\}$ and $\{\mu_{u,k,k',v}\}$ requires $\mathcal{O}(K^2)$ and $\mathcal{O}(K^2V^2)$ time, respectively. Estimating $\{\beta_{kk'}\}$ and $\{\psi_{u,k,k',v}\}$ requires solving a linear system, and takes $\mathcal{O}(K^2D^3)$ and $\mathcal{O}(K^2\bar{N})$ time, where $\bar{N}$ denotes the number of node pairs with at least one interaction in $[0, T]$. To sample the Pólya-Gamma variables $\{\omega_{u,k,k',v}\}$, we employed a fast and accurate approximate sampler of Zhou (M. Zhou 2016), which matches the first two moments of the original distribution. Using the EM algorithm, the Pólya-Gamma variables are updated in closed-form (as a hyperbolic function) (Scott and Sun 2013).

Algorithm 3 Gibbs Sampler for the Hawkes Edge Partition Model

```
Require: events data $\mathcal{D} = \{(t_i, s_i, d_i)\}_{i=1}^{N}$, $\{\Phi, \Omega\}$ inferred by the HGP-EPM, maximum iterations $\mathcal{J}$
Ensure: $\{\mu_{u,k,k',v}\}$, $\{\alpha_{kk'}\}$, $\{(z_i^s, z_i^d)\}$
for $l = 1{:}\mathcal{J}$ do
    for $n = 1{:}N$ do
        Sample $b_i$ (Eq. 4.4.1)
        Sample the latent variables $(z_i^s, z_i^d)$ (Eq. 4.4.2)
        Update the intensity function $\lambda_{d_i, z_i^d, z_i^s, s_i}(t_i)$ (Eq. 4.4.3)
    end for
    Update $m^{\text{exo}}_{u,k,k',v}$ and $m^{\text{endo}}_{u,k,k',v}$ (Eq. 4.4.4)
    Sample the base intensities $\{\mu_{u,k,k',v}\}$ (Eq. 4.4.7)
    Sample the parameters $\{\beta_{kk'}\}$, $\{\omega_{u,k,k',v}\}$, $\{\psi_{u,k,k',v}\}$ (Eqs. 4.4.10; 4.4.8)
    Sample the kernel parameters $\{\alpha_{k,k'}\}$ (Eq. 4.4.6)
end for
```

Algorithm 4 Expectation-Maximization algorithm for the Hawkes Edge Partition Model

```
Require: events data $\mathcal{D} = \{(t_i, s_i, d_i)\}_{i=1}^{N}$, $\{\Phi, \Omega\}$ inferred by the HGP-EPM
Ensure: $\{\mu_{u,k,k',v}\}$, $\{\alpha_{kk'}\}$
repeat
    for $n = 1{:}N$ do
        Update $(p^{\text{exo}}_{ikk'}, p^{\text{endo}}_{ikk'})$ (Eq. 4.4.11)
        Update the intensity function $\lambda_{d_i, s_i}(t_i)$
    end for
    Update $m^{\text{exo}}_{u,k,k',v}$ and $m^{\text{endo}}_{u,k,k',v}$ (Eq. 4.4.12)
    Update the base intensities $\{\mu_{u,k,k',v}\}$ (Eq. 4.4.15)
    Update the parameters $\{\beta_{kk'}\}$, $\{\omega_{u,k,k',v}\}$, $\{\psi_{u,k,k',v}\}$ (Eqs. 4.4.18 4.4.14; 4.4.17)
    Update the kernel parameters $\{\alpha_{k,k'}\}$ (Eq. 4.4.16)
until convergence
```

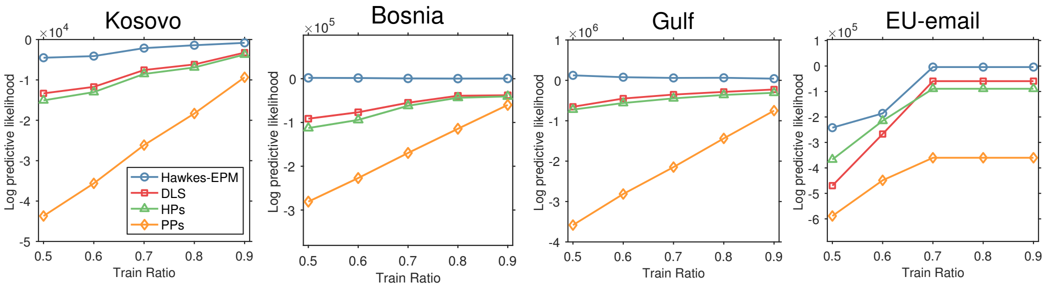


Figure 4.3: The predictive log-likelihood of each method in four benchmark datasets. The results are averaged over ten runs.

4.5 EXPERIMENTS

We evaluate the proposed Hawkes-EPM model on four benchmark temporal interaction events datasets, Kosovo, Bosnia, Gulf, EU-email: (i) **Kosovo.** This dataset[2] consists of the interaction events among 168 nations over 451 days (04/01/1998-31/03/1999). There are 1139 edges and 7224 interactions. We utilized the auxiliary events attributes (e.g., military force or support) aggregated over the whole time interval between any two nodes as their covariate data. (ii) **Bosnia.** This dataset consists of the interaction events among 159 nations over 1819 days (17/01/1991-31/12/1995). There are 1918 edges, and 34014 interactions. (iii) **Gulf.** This dataset[3] consists of the 304401 interaction events among 202 nations over 7291 days (15/04/1979-31/03/1999). There are 7184 edges. (iv) **EU-email.** This dataset[4] consists of the 332334 email communications among 1005 individuals over 526 days. There are 24929 edges. We generated the covariate data between each pair of nodes using their common attributes.

We compared our model to (i) a Poisson process (PPs) model, which independently models the interaction dynamics between each pair of nodes by a constant event rate, (ii) a Hawkes process (HPs) model, in which we assume the same base rate and kernel parameters for each pair of nodes. Following (J. Yang et al. 2017), we utilized four basis kernels–three exponential kernels with time decaying scale: one hour, one day, one week respectively: $\gamma^1(t) \equiv \exp(-24t), \gamma^2(t) \equiv \exp(-t), \gamma^3(t) \equiv \exp(-t/7)$, and a periodic kernel $\gamma^4(t) \equiv \exp(-t/7)\sin^2(\pi t/7)$, (iii) the Hawkes Dual Latent Space (DLS) model (J. Yang et al. 2017)[5], which captures the base event rate with the Latent space model (P. D. Hoff et al. 2001), and models the reciprocating dynamics in each particular interaction. Given the aggregated graph, we estimated the parameters $\{\boldsymbol{\Phi}, \boldsymbol{\Lambda}\}$ of the HGP-EPM with the truncation level $K_{\max} = 100$. We ran the Gibbs sampler detailed in (M. Zhou et al. 2015) for 10000 MCMC iterations, and used the maximum a posterior estimate $\{\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Lambda}}\}$ in the second step. For the Hawkes-EPM, we found that estimating δ suffered from identifiability issues as reported in (J. Yang et al. 2017; Tan, Rao, and Neville 2018; Tan et al. 2018), and choose a kernel decay of $\delta = 1/10$. In our experiments, both the Gibbs sampler and the EM algorithm perform comparably, and we only report the results obtained with EM algorithm.

---

[2] `http://eventdata.parusanalytics.com/data.dir/pevehouse.html`.

[3] `http://eventdata.parusanalytics.com/data.dir/gulf.html`.

[4] `http://snap.stanford.edu/data/email-EuAll.html`.

[5] `https://github.com/jiaseny/lspp`.

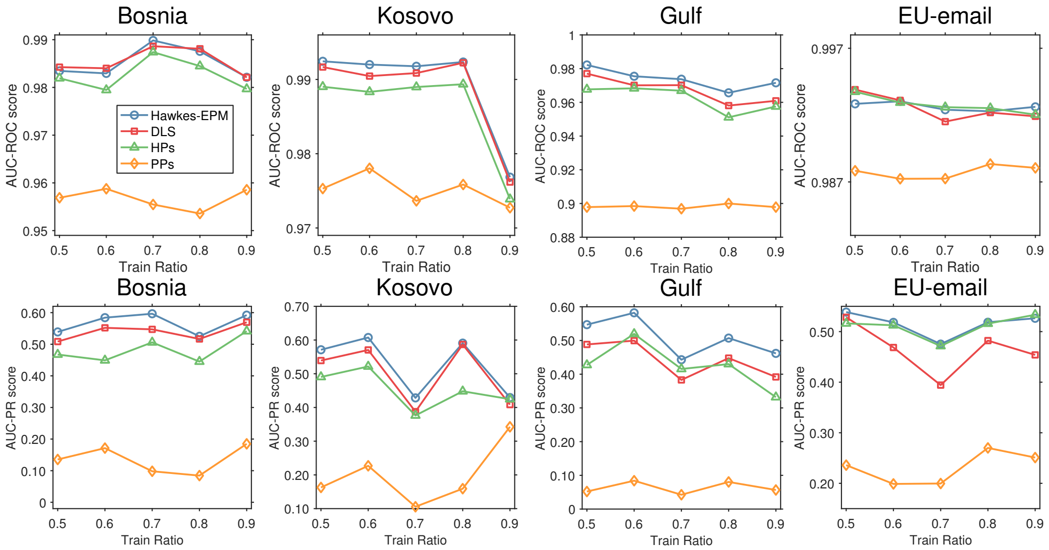


Figure 4.4: AUC-ROC and PR scores for the temporal link prediction.

4.5.1 *Predictive log-likelihood.*

To evaluate the predictive performance, we sorted the interaction events according to the corresponding timestamps, and made a train-test split so that the training dataset consists of 50%-90% of the whole events. We trained all the methods on the training data, and calculated their predictive log-likelihood over the test dataset. In Figure (4.3), we report the average log-likelihood of each method applied to four benchmark datasets over ten runs. Overall, we found that the Hawkes process based models (HPs, DLS, the Hawkes-EPM) significantly outperform the Poisson process model, which confirms that most interactions arised as responses to the past events of their opposite directions (reciprocity). We also found that DLS performs slightly better than the Hawkes process model although DLS accounts for heterogeneity both in the base rate for each pair of nodes, and also in their each specific interaction. Overall, the Hawkes-EPM achieves the higher predictive log-likelihood compared with DLS and the Hawkes process model. We conjecture that this is because most entities have very few interactions. For most entities exhibiting few interactions in training dataset, DLS fails to accurately capture their interaction dynamics by accounting for each particular interaction. The Hawkes EPM allows the latent patterns to be shared among entities with similar latent features, and thus captures the interaction dynamics of those entities exhibiting few interactions more accurately.

4.5.2 *Temporal link prediction.*

We trained all the methods using the training datasets as we used in calculating predictive log-likelihood. In this task, we let all the models to predict the probability that an edge appears (at least one interaction occurrs) between each pair of nodes in the time interval $[t, t+\hat{\pi})$ with t being the end time of the training events. We also set $\hat{\pi}$ to be 50 days for all the datasets. We calculated the probability of there being at least one interaction in $[t, t+\hat{\pi})$ as $1 - \exp\{-\int_t^{t+\hat{\pi}} \lambda_{uv}(s)\mathrm{d}N_{vu}(s)\}$. Finally, we computed the average area under the curve (AUC) of both the receiver operating characteristic (ROC) and precision-recall (PR) to evaluate the predictive performance. As shown in Figure 4.4, the Hawkes process based models (HPs, DLS, Hawkes-EPM) capture the reciprocating dynamics of the interactions among nodes, and thus significantly outperform the Poisson process model. We note that AUC-PR is more sensitive to the proportion of true edges in the top ranked ones, and also noticed that most node pairs exhibit no edges in the time interval $[t, t+\hat{\pi})$. A closer looking into the AUC-PR scores, we found that the Hawkes-EPM performs better than HPs and DLS when the training ratio is low. This is because the Hawkes-EPM shares the kernel parameters among node pairs, and thus performs well even if most node pairs exhibit few interactions.

4.5.3 *Exploratory analysis.*

We also used the Gulf dataset to explore the latent structure estimated by the Hawkes-EPM. We found that $K = 12$ latent communities, and most of those communities correspond to international military conflicts among nations. Figs. 4.5 to 4.7 show the inferred intensities of the interaction among these nations. For instance, we found that the peaks of the intensities bewteen USA - Iraq(IRQ) correspond to events surrounding the Gulf War (1990-1991), the Cruise missile attack on Iraq in 1993 and 1996, the Bombing of Iraq in 1998. In addition, we also plot the intensities of interaction events between Iran(IRN)-Iraq(IRQ). The intensities of the interaction events between these two nodes are gradually increasing from 1980, and reach the peak at 1988. To interpret the inferred interaction dynamics between these two nodes, we performed a web search, and found that the Iran-Iraq War started on September, 1980 and ended on August, 1988. Most of the inferred intensities between each pair of nations in the Gulf dataset confirm our knowledge of international affairs.

4.6 CONCLUSIONS

We presented a probabilistic framework, the Hawkes edge partition model (Hawkes-EPM) for inferring the implicit community structure and reciprocating dynamics among entities from their event-based temporal interactions. The Hawkes-EPM not only models the inherent overlapping communities, sparsity and degree heterogeneity behind the observed interactions, but also captures how the latent communities influence the interaction dynamics among their involved entities. Experimental results demonstrate the interpretability and competitive predictive performance of our model in several real-world datasets.

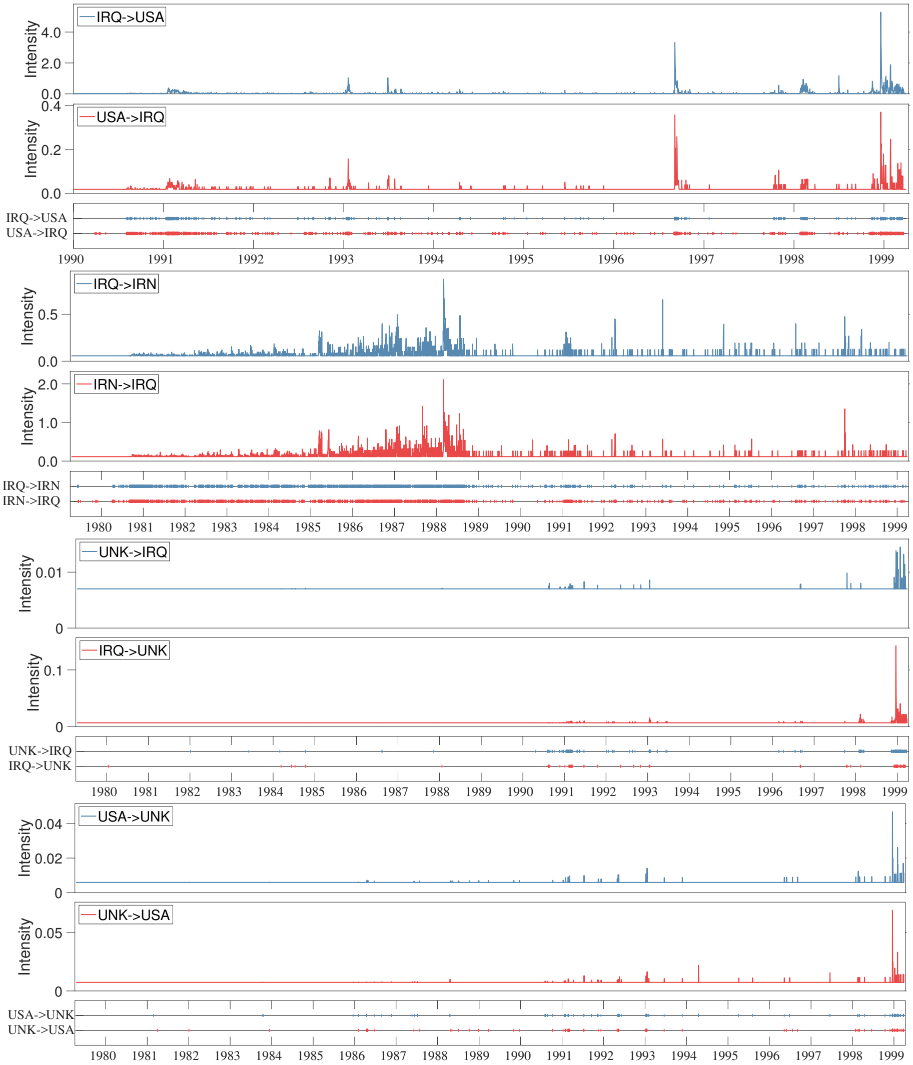


Figure 4.5: The plots show the intensity of interaction events among nations inferred by the Hawkes-EPM in the Gulf dataset.

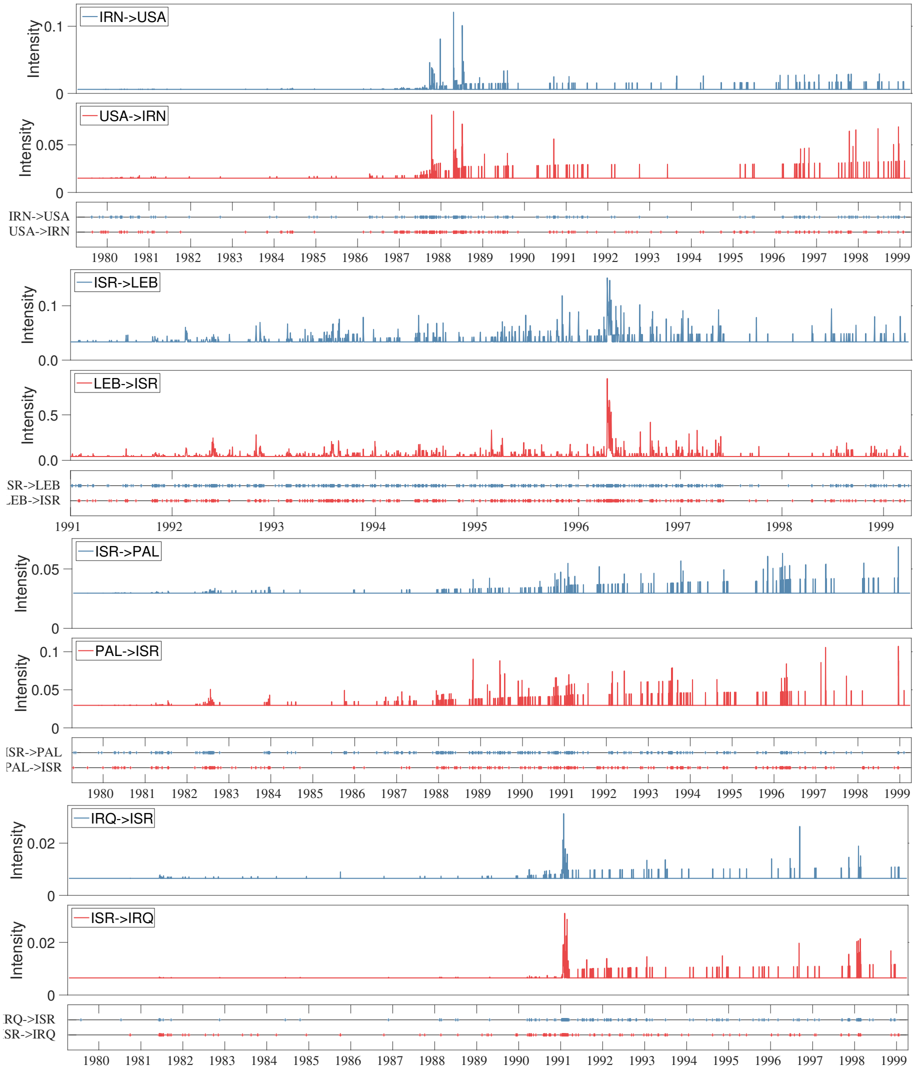


Figure 4.6: The plots show the intensity of interaction events among nations inferred by the Hawkes-EPM in the Gulf dataset.

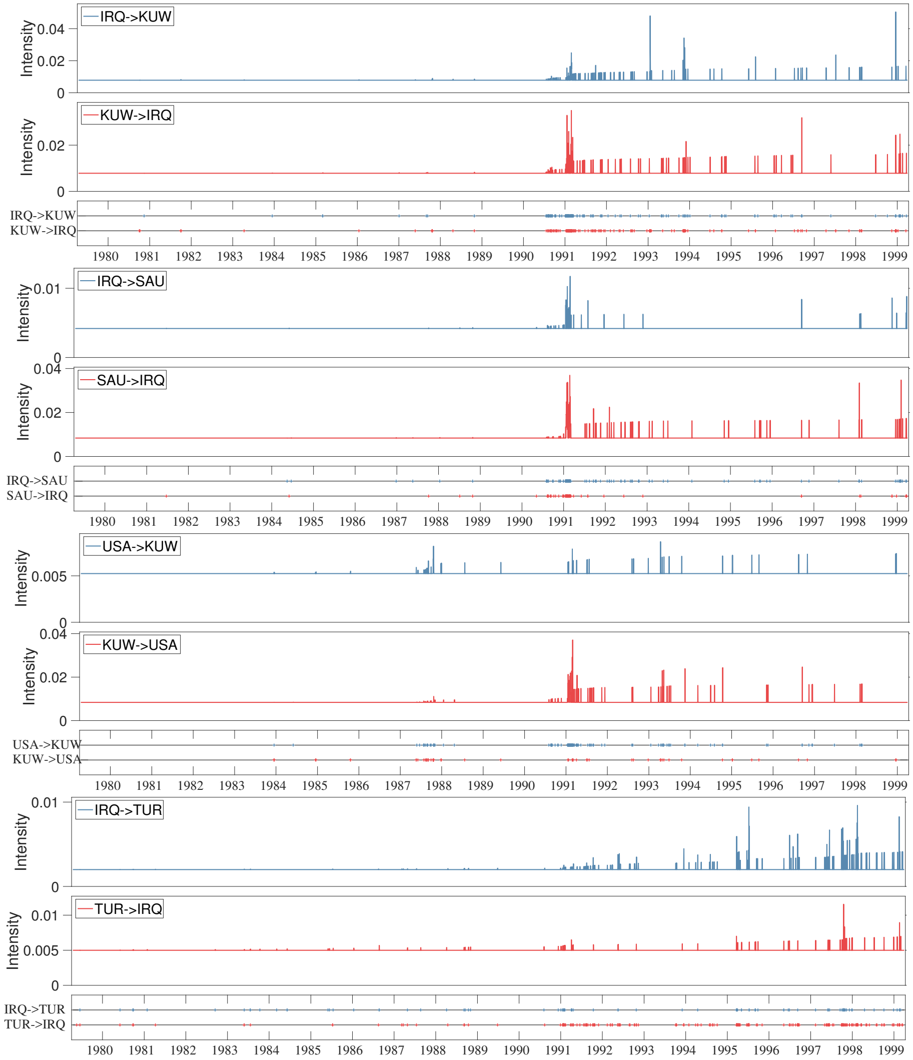


Figure 4.7: The plots show the intensity of interaction events among nations inferred by the Hawkes-EPM in the Gulf dataset.

5

STOCHASTIC GRADIENT MARKOV CHAIN MONTE CARLO FOR DISCRETE TIME NETWORK MODELS

In the earlier chapters, we have demonstrated the successful applications of the Poisson gamma memberships models for overlapping community detection and missing edge prediction in dynamic networks. In this chapter, we propose a novel generative model that extends the Poisson gamma memberships framework to model temporal assortative graphs by endowing each node with a positive memberships vector, constructed using Dirichlet prior specification, which retains the expressiveness and interpretability of the original gamma process edge partition model (GaP-EPM). Specifically, the new model utilizes a Dirichlet Markov chain to capture the smooth evolution of the nodes' memberships over time. In particular, the unique construction of the Dirichlet Markov chain enables us to adopt the recently advanced SG-MCMC algorithms (Patterson et al. 2013; T. Chen et al. 2014; Ding et al. 2014; Ma et al. 2015; C. Chen et al. 2016) for scalable and parallelizable inference in the proposed model. The remainder of the chapter is structured as follows. We first review relevant background. Then, we present the novel dynamic edge partition model, and describe its Gibbs sampler and stochastic gradient Markov chain Monte Carlo algorithm. The accuracy and efficiency of our method are demonstrated on several real-world datasets. Finally, we conclude the chapter.

5.1 STOCHASTIC GRADIENT MARKOV CHAIN MONTE CARLO

Stochastic gradient Markov chain Monte Carlo (SG-MCMC) is an approximate MCMC algorithm that subsamples the data, and uses the stochastic gradients to update the parameters of interest at each step. Given a dataset $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$, we have a generative model $p(X \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is drawn from the prior $p(\boldsymbol{\theta})$. Our aim is to compute the posterior of $\boldsymbol{\theta}$, i.e., $p(\boldsymbol{\theta} \mid X) \propto \exp(-H(\theta))$ with potential function $H(\boldsymbol{\theta}) \equiv -\sum_{\boldsymbol{x}_i \in X} \log p(\boldsymbol{x}_i \mid \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})$. It has been shown (Ma et al. 2015) that $p^s(\boldsymbol{\theta}) \propto \exp(-H(\boldsymbol{\theta}))$ is a stationary distribution of the dynamics of a stochastic differential equation of the form as

$$\mathrm{d}\boldsymbol{\theta} = f(\boldsymbol{\theta})\mathrm{d}t + \sqrt{2\mathbf{D}(\boldsymbol{\theta})}\mathrm{d}\mathbf{W}(t),$$

if $f(\boldsymbol{\theta})$ is restricted to the following form as

$$\begin{aligned} f(\boldsymbol{\theta}) &= [\mathbf{D}(\boldsymbol{\theta}) + \mathbf{Q}(\boldsymbol{\theta})]\nabla H(\boldsymbol{\theta}) + \Gamma(\boldsymbol{\theta}), \\ \Gamma_i(\boldsymbol{\theta}) &= \sum_{j=1}^{d} \frac{\partial}{\partial \boldsymbol{\theta}_j} \left(\mathbf{D}_{ij}(\boldsymbol{\theta}) + \mathbf{Q}_{ij}(\boldsymbol{\theta})\right), \end{aligned} \tag{5.1.1}$$

where $f(\boldsymbol{\theta})$ is the deterministic drift, $\mathbf{W}(t)$ is d–dimensional Wiener process, $\mathbf{D}(\boldsymbol{\theta})$ is a positive semi-definite diffusion matrix, and $\mathbf{Q}(\boldsymbol{\theta})$ is skew-symmetric curl matrix. This leads to the following update rule used in SG-MCMC algorithms as

$$\begin{aligned}\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \epsilon_t[(\mathbf{D}(\boldsymbol{\theta}_t) + \mathbf{Q}(\boldsymbol{\theta}_t))\nabla \tilde{H}(\boldsymbol{\theta}_t) + \Gamma(\boldsymbol{\theta}_t)]& \\ + \mathcal{N}(\mathbf{0}, \epsilon_t(2\mathbf{D}(\boldsymbol{\theta}_t) - \epsilon_t\hat{\mathbf{B}}_t)),& \\ \tilde{H}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) - \rho \sum_{\boldsymbol{x}_i \in \tilde{X}} \log p(\boldsymbol{x}_i \mid \boldsymbol{\theta}),&\end{aligned} \tag{5.1.2}$$

where $\{\epsilon_t\}$ is a sequence of step sizes, $\tilde{X}$ is the mini-batch subsampled from the full data X, $\rho \equiv |X|/|\tilde{X}|$, and $\hat{\mathbf{B}}_t$ is the estimate of stochastic gradient noise variance.

As shown in (Ma et al. 2015), setting $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})^{-1}$ where $\mathbf{G}(\boldsymbol{\theta})$ is the Fisher information matrix, and $\mathbf{Q}(\boldsymbol{\theta}) = 0$ in Eq.(5.1.2), we obtain the update rule of the stochastic gradient Riemannian Langevin dynamics (SGRLD) as

$$\begin{aligned}\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \epsilon_t[(\mathbf{G}(\boldsymbol{\theta}_t)^{-1}\nabla \tilde{H}(\boldsymbol{\theta}_t) + \Gamma(\boldsymbol{\theta}_t)]& \\ + \mathcal{N}\left(\mathbf{0}, 2\epsilon_t \mathbf{G}(\boldsymbol{\theta}_t)^{-1}\right).&\end{aligned} \tag{5.1.3}$$

5.2 THE DYNAMIC DIRICHLET EDGE PARITITION MODEL

As we introduced in previous chapters, let $\{A^{(t)}\}_{t=1}^{T}$ be a sequence of networks or graphs, where $A^{(t)} \in \{0,1\}^{V\times V}$ is the network snapshot observed at time t with V being the number of nodes. An edge is present between nodes u and v, i.e., $A_{uv}^{(t)} = 1$ if they are connected at time t. Otherwise, $A_{uv}^{(t)} = 0$. We ignore self edges $A_{uu}^{(t)}$. Generally, the considered temporal network can be decomposed into a set of K communities, where K is generally unknown *a priori*. In order to extract an overlapping community structure from the given network, we represent each node u at time t by a K–dimensional positive memberships vector $\{\phi_{uk}^{(t)}\}_{k=1}^{K}$, and thus each of the K memberships can be considered as how *actively* it is involved in the corresponding community at that time. In temporal networks, the observed edges among nodes change over time because the association relationships of these nodes to the underlying communities are evolving (J. R. Foulds et al. 2011; Heaukulani et al. 2013; M. Kim et al. 2013). Hence, learning an expressive and interpretable nodes representations is the key to understanding the true dynamics of the underlying relations between nodes. Unlike most existing methods (J. R. Foulds et al. 2011; Heaukulani et al. 2013; M. Kim et al. 2013) utilizing a factorial hidden Markov model to capture the evolution of *binary* nodes' memberships, we use a Dirichlet Markov chain construction to allow node-community memberships to vary smoothly over time. More specifically, for each active community k, we draw $\boldsymbol{\phi}_k^{(t)}$ from a Dirichlet distribution, i.e., $\{\phi_{uk}^{(t)}\}_{u=1}^{V} \sim \text{Dirichlet}(\eta V\phi_{1k}^{(t-1)}, \ldots, \eta V\phi_{Vk}^{(t-1)})$, for $t \in \{2, \ldots, T\}$, where $\phi_{uk}^{(t)}$ corresponds to the membership of node u to community k at time t. In particular, we draw $\boldsymbol{\phi}_k^{(1)} \sim \text{Dirichlet}(\boldsymbol{\eta})$. The intuition behind this construction is that each community can be thought of as a distribution over the V nodes (akin to a topic model). In temporal networks, these communities are evolving over time because the memberships of their affiliated nodes are varying. Moreover, for each community k, we draw an associated weight $r_k \sim \text{Gamma}(g_k, p_k/(1-p_k))$, where $p_k \sim \text{Beta}(c_0\alpha, c_0(1-\alpha))$, to modulate the interactions probability between any two nodes affiliated to that community. Note that the hierarchical beta-gamma prior for r_k allows inferring the appropriate number of latent communities by shrinking the redundant community weights to zeros (M. Zhou and L. Carin 2015b). Finally, an edge between each pair of nodes is generated using the Bernoulli-Poisson link function (Dunson et al. 2005; M. Zhou et al. 2015) as

$$\begin{aligned} A_{uv}^{(t)} &\sim \mathbb{1}(\tilde{A}_{uv}^{(t)} \geq 1), \\ \tilde{A}_{uv}^{(t)} &\sim \text{Poisson}\Big(\sum_{k=1}^{K} \phi_{uk}^{(t)} r_k \phi_{vk}^{(t)}\Big), \end{aligned}$$

where $\phi_{uk}^{(t)} r_k \phi_{vk}^{(t)}$ measures how *strongly* nodes u and v are connected at time t because they are both affiliated to community k. Hence, naturally, the probability that a pair of nodes are connected at time t, will be higher if the two nodes share more common communities at that time. Note that sampling of $\{\tilde{A}_{uv}^{(t)}\}_{u,v,t}$ only needs to be performed using rejection sampler (M. Zhou et al. 2015) on nonzero entries in a given network as

$$(\tilde{A}_{uv}^{(t)} \mid -) \sim \begin{cases} \delta(0), & \text{if } A_{uv}^{(t)} = 0 \\ \text{Poisson}_{+}\Big(\sum_{k=1}^{K} \phi_{uk}^{(t)} r_k \phi_{vk}^{(t)}\Big), & \text{otherwise} \end{cases} \tag{5.2.1}$$

where $\delta(0)$ is a point measure concentrated at 0, Poisson_{+} is the zero-truncated Poisson distribution with support only on the positive integers, and "–" represents all other variables. Hence, inference in this model scales linearly with the number of nonzero edges in the given network data.

The full generative construction of the proposed model is as follows:

$$\begin{aligned} r_k &\sim \text{Gamma}(g_k, p_k/(1-p_k)), \\ p_k &\sim \text{Beta}(c_0\alpha, c_0(1-\alpha)), \\ \{\phi_{uk}^{(t)}\}_{u=1}^{V} &\sim \text{Dirichlet}\left(\eta V \{\phi_{uk}^{(t-1)}\}_{u=1}^{V}\right), \text{ for } t \in \{2, \ldots, T\} \\ \{\phi_{uk}^{(1)}\}_{u=1}^{V} &\sim \text{Dirichlet}(\eta \mathbf{1}_N), \\ \eta &\sim \text{Gamma}(a_0, 1/b_0), \\ \tilde{A}_{uv}^{(t)} &\sim \text{Poisson}\Big(\sum_{k=1}^{K} \phi_{uk}^{(t)} r_k \phi_{vk}^{(t)}\Big), \\ A_{uv}^{(t)} &\sim \mathbb{1}(\tilde{A}_{uv}^{(t)} \geq 1). \end{aligned}$$

5.3 INFERENCE

Despite the proposed model not being natively conjugate for exact inference, we leverage the Negative-Binomial augmentation technique to derive a simple-to-implement Gibbs sampler with closed-form update equations. For large temporal network data, we develop stochastic gradient MCMC algorithms using both the expanded-mean and reduced-mean re-parameterization tricks.

5.3.1 *Batch Gibbs Sampler*

We now proceed to describe our batch Gibbs sampler for the proposed model. The parameters that need to be inferred are $\{\tilde{A}_{uv}^{(t)}\}_{u,v,t}$, $\{\tilde{A}_{uvk}^{(t)}\}_{u,v,k,t}$, $\{\phi_{uk}^{(t)}\}_{u,k,t}$, $\{r_k\}_k$, $\{p_k\}_k$ and η.

We present the inference update equations for each of the parameters below.

Sampling of $\{\tilde{A}^{(t)}_{uvk}\}_{u,v,t}$: Using the Poisson-multinomial equivalence (M. Zhou and L. Carin 2015b), the latent counts $m^{(t)}_{uvk}$ are sampled from a multinomial distribution as

$$\left(\{\tilde{A}^{(t)}_{uvk}\}_{k=1}^{K} \mid -\right) \sim \text{Multinomial}\left(\tilde{A}^{(t)}_{uv}, \frac{\{\phi^{(t)}_{uk} r_k \phi^{(t)}_{vk}\}_{k=1}^{K}}{\sum_{k=1}^{K} \phi^{(t)}_{uk} r_k \phi^{(t)}_{vk}}\right). \tag{5.3.1}$$

Sampling of $\{\phi^{(t)}_{uk}\}_{u,k,t}$: According to the additive property of the Poisson distribution, we have the aggregated counts $\tilde{A}^{(t)}_{u\cdot k} \equiv \sum_{v \neq u} m^{(t)}_{uvk}$ and $\tilde{A}_{\cdot\cdot k} \equiv \frac{1}{2}\sum_{u,v,t} m^{(t)}_{uvk}$ that can be expressed as

$$\tilde{A}^{(t)}_{u\cdot k} \sim \text{Poisson}(r_k \phi^{(t)}_{uk}), \tag{5.3.2}$$

$$\tilde{A}_{\cdot\cdot k} \sim \text{Poisson}(r_k T). \tag{5.3.3}$$

Via the Poisson-multinomial equivalence, we can equivalently sample $\{\tilde{A}^{(t)}_{u\cdot k}\}_{u=1}^{V}$ as

$$\{\tilde{A}^{(t)}_{u\cdot k}\}_{u=1}^{V} \sim \text{Multinomial}\left(\tilde{A}^{(t)}_{\cdot\cdot k}, \{\phi^{(t)}_{uk}\}_{u=1}^{V}\right). \tag{5.3.4}$$

For $t = T$, we sample the Dirichlet distributed vector $\{\phi^{(t)}_{uk}\}_{u=1}^{V}$ using the Dirichlet-multinomial conjugacy as

$$(\{\phi^{(t)}_{uk}\}_{u=1}^{V} \mid -) \sim \text{Dirichlet}\left(\eta V \{\phi^{(t-1)}_{uk} + \tilde{A}^{(t-1)}_{u\cdot k}\}_{u=1}^{V}\right). \tag{5.3.5}$$

For $2 \leq t \leq (T-1)$, as we already have the multinomial likelihood and Dirichlet prior as

$$\{\tilde{A}^{(t+1)}_{u\cdot k}\}_{u=1}^{V} \sim \text{Multinomial}\left(\tilde{A}^{(t+1)}_{\cdot\cdot k}, \{\phi^{(t+1)}_{uk}\}_{u=1}^{V}\right), \tag{5.3.6}$$

$$\{\phi^{(t+1)}_{uk}\}_{u=1}^{V} \sim \text{Dirichlet}\left(\eta V \{\phi^{(t)}_{uk}\}_{u=1}^{V}\right). \tag{5.3.7}$$

Marginalizing out $\{\phi^{(t+1)}_{uk}\}_{u=1}^{V}$ using the Dirichlet-multinomial conjugacy leads to

$$\{\tilde{A}^{(t+1)}_{u\cdot k}\}_{u=1}^{V} \sim \text{DirMult}\left(\tilde{A}^{(t+1)}_{\cdot\cdot k}, \eta V \{\phi^{(t)}_{uk}\}_{u=1}^{V}\right). \tag{5.3.8}$$

Using the Negative-Binomial augmentation strategy, we introduce an auxiliary variable

$$\zeta^{(t+1)}_{k} \sim \text{Beta}(\tilde{A}^{(t+1)}_{\cdot\cdot k}, \eta V), \tag{5.3.9}$$

and then the latent counts $\{\tilde{A}^{(t+1)}_{u\cdot k}\}_{u=1}^{V}$ can be equivalently sampled as

$$\tilde{A}^{(t+1)}_{u\cdot k} \sim \text{NB}\left(\eta V \phi^{(t)}_{uk}, \zeta^{(t+1)}_{k}\right).$$

We further augment $\tilde{A}^{(t+1)}_{u\cdot k}$ with an auxiliary CRT distributed variable as

$$\xi^{(t+1)}_{uk} \sim \text{CRT}(\tilde{A}^{(t+1)}_{u\cdot k}, \eta V \phi^{(t)}_{uk}). \tag{5.3.10}$$

According to the Poisson Logarithmic bivariate distribution, we can equivalently draw $\tilde{A}^{(t+1)}_{u\cdot k}$ and $\xi^{(t+1)}_{uk}$ as

$$\tilde{A}^{(t+1)}_{u\cdot k} \sim \text{SumLog}\left(\xi^{(t+1)}_{uk}, \zeta^{(t+1)}_{k}\right),$$

$$\xi^{(t+1)}_{uk} \sim \text{Poisson}\left[-\eta V \phi^{(t)}_{uk} \log\left(1 - \zeta^{(t+1)}_{k}\right)\right].$$

Via the Poisson-multinomial equivalence, we can sample $\{\xi_{uk}^{(t+1)}\}_{u=1}^{N}$ from a multinomial distribution as

$$\{\xi_{uk}^{(t+1)}\}_{u=1}^{V} \sim \text{Multinomial}\left(\xi_{\cdot k}^{(t+1)}, \{\phi_{uk}^{(t)}\}_{u=1}^{V}\right). \tag{5.3.11}$$

Combining the prior placed over $\{\phi_{uk}^{(t)}\}_{u=1}^{V}$ and the multinomial likelihood function in Eqs.(5.3.4;5.3.11), we sample $\{\phi_{uk}^{(t)}\}_{u=1}^{V}$ using the Dirichlet-multinomial conjugacy as

$$(\{\phi_{uk}^{(t)}\}_{u=1}^{V} \mid -) \sim \text{Dirichlet}\left(\{\eta V \phi_{uk}^{(t-1)} + \xi_{uk}^{(t+1)} + \tilde{A}_{u\cdot k}^{(t)}\}_{u=1}^{V}\right), \tag{5.3.12}$$

where $\xi_{uk}^{(t+1)}$ can be considered as the information passed back from time $t+1$ to t.

Sampling of η: As we already have the Poisson likelihood $\xi_{uk}^{(t)} \sim \text{Poisson}(-\eta V \phi_{uk}^{(t-1)} \log(1 - \zeta_k^{(t)}))$ and the gamma prior $\eta \sim \text{Gamma}(a_0, 1/b_0)$, we sample η using the gamma-Poisson conjugacy as

$$(\eta \mid -) \sim \text{Gamma}\left(a_0 + \sum_{u,k,t} \xi_{uk}^{(t)}, \frac{1}{b_0 - V\sum_{k,t}[\log(1-\zeta_k^{(t)})]}\right). \tag{5.3.13}$$

Sampling of r_k: Similaly, using the gamma-Poisson conjugacy, we obtain the conditional distribution of r_k as

$$(r_k \mid -) \sim \text{Gamma}\left(g_k + \tilde{A}_{\cdot\cdot k}, \frac{p_k}{1 + (T-1)p_k}\right). \tag{5.3.14}$$

Sampling of p_k: Marginalizing out r_k from the likelihood in Eq.(5.3.3) and the prior $r_k \sim \text{Gamma}(g_k, p_k/(1-p_k))$, we obtain $\tilde{A}_{\cdot\cdot k}/T \sim \text{NB}(g_k, p_k)$.Using the beta-negative-binomial conjugacy, we sample p_k as

$$(p_k \mid -) \sim \text{Beta}\left(c_0\alpha + \tilde{A}_{\cdot\cdot k}/T, c_0(1-\alpha) + g_k\right). \tag{5.3.15}$$

The full inference procedure is presented in Algorithm 5.

5.3.2 *Scalable Inference via Stochastic Gradient MCMC*

While the proposed Gibbs sampler scales linearly with the number of nonzero entries in the given temporal network data, Gibbs sampler tends to be slow to mix and converge in practice. In order to mitigate this limitation, we resort to SG-MCMC algorithms for scalable inference in the proposed model. Our SG-MCMC algorithm for the proposed model is mainly based on the stochastic gradient Riemannian Langevin dynamics for the probability simplex because of the unique construction of the Dirichlet Markov chain here. Naively applying SG-MCMC to perform inference for the probability simplex may result in invalid values being proposed. Thus, various strategies have been investigated to parameterize the probability simplex (Patterson et al. 2013).

First, we consider the expanded-mean that was shown to achieve overall best performance (Patterson et al. 2013). In the proposed model, $\boldsymbol{\phi}_k^{(t)}$ is an V–dimensional probability simplex, and our goal is to update $\boldsymbol{\phi}_k^{(t)}$ as the global parameter on a mini-batch data at each step. Using the expanded-mean trick, we parameterize $\boldsymbol{\phi}_k^{(t)}$ as

Algorithm 5 Batch Gibbs Sampler for the proposed Dynamic Dirichlet Edge Partition Model

```
Require: temporal graphs $\{A^{(t)}\}_t$, maximum iterations $\mathcal{J}$
Ensure: posterior mean $\{\boldsymbol{\phi}^{(t)}_k\}_{k,t}, \{r_k\}_k, \{p_k\}_k, \eta$
for $l = 1{:}\mathcal{J}$ do
    Sample $\{\tilde{A}^{(t)}_{uv}\}_{u,v,t}$ and $\{\tilde{A}^{(t)}_{uvk}\}_{u,v,k,t}$ (Eqs. 5.2.1; 5.3.1)
    Update $\{\tilde{A}^{(t)}_{\cdot\cdot k}\}_{k,t}, \{\tilde{A}^{(t)}_{\cdot\cdot k}\}_{k,t}$, and $\{\tilde{A}_{\cdot\cdot k}\}_k$
    for t = T,..., 1 do
        Sample $\{\boldsymbol{\zeta}^{(t)}_k\}_k$ and $\{\zeta^{(t)}_k\}_k$ (Eqs. 5.3.10; 5.3.9)
    end for
    for t = 1,..., T do
        Sample $\{\boldsymbol{\phi}^{(t)}_k\}_k$ (Eqs. 5.3.5; 5.3.12)
    end for
    Sample $\{r_k\}_k, \{p_k\}_k$ and $\eta$ (Eqs. 5.3.14; 5.3.15; 5.3.13)
end for
```

$\{\phi^{(t)}_{1k}, \ldots, \phi^{(t)}_{Vk}\} = \{\hat{\phi}^{(t)}_{1k}, \ldots, \hat{\phi}^{(t)}_{Vk}\} / \hat{\phi}^{(t)}_{\cdot k}$ where $\hat{\phi}^{(t)}_{uk} \sim \mathrm{Gamma}(\eta V \phi^{(t-1)}_{uk}, 1)$ and $\hat{\phi}^{(t)}_{\cdot k} \equiv \sum_i \hat{\phi}^{(t)}_{uk}$. Then, $\{\hat{\phi}^{(t)}_{1k}, \ldots, \hat{\phi}^{(t)}_{Vk}\} / \hat{\phi}^{(t)}_{\cdot k}$ will follow Dirichlet $\left(\eta V \{\phi^{(t-1)}_{uk}\}_{u=1}^{V}\right)$ distribution.

Given the log-posterior of $\hat{\boldsymbol{\phi}}^{(t)}_k$ on the full data A as

$$\log p(\{\hat{\phi}^{(t)}_{uk}\}_{u=1}^{V} \mid -) \propto \sum_{u=1}^{V} \left[(\tilde{m}^{(t)}_{uk} + \eta V \phi^{(t-1)}_{uk} - 1) \log(\hat{\phi}^{(t)}_{uk}) + \tilde{m}^{(t)}_{uk} \log(\hat{\phi}_{\cdot k}) - \hat{\phi}^{(t)}_{uk} \right]$$

where $\tilde{m}^{(t)}_{uk} \equiv \zeta^{(t+1)}_{uk} + \tilde{A}^{(t)}_{u\cdot k}$, we take the gradient of the log-posterior with respect to $\hat{\boldsymbol{\phi}}^{(t)}_k$ on a mini-batch data $\hat{A}$, and then obtain

$$\nabla_{\hat{\boldsymbol{\phi}}^{(t)}_k} [-\tilde{H}(\hat{\boldsymbol{\phi}}^{(t)}_k)] = \left\{ \frac{\rho \tilde{m}^{(t)}_{uk} + \eta V \phi^{(t-1)}_{uk} - 1}{\hat{\phi}^{(t)}_{uk}} - \frac{\rho \tilde{m}^{(t)}_{\cdot k}}{\hat{\phi}^{(t)}_{\cdot k}} - 1 \right\}_{u=1}^{N}, \tag{5.3.16}$$

where $\rho \equiv |A| / |\hat{A}|$, and $\tilde{m}^{(t)}_{\cdot k} \equiv \zeta^{(t+1)}_{\cdot k} + \tilde{A}^{(t)}_{\cdot\cdot k}$.

Given the gamma-Poisson construction used in expanded mean $\tilde{m}^{(t)}_{uk} \sim \mathrm{Poisson}(\hat{\phi}^{(t)}_{uk})$, the Fisher information matrix is calculated as

$$\mathbf{G}\left(\hat{\boldsymbol{\phi}}^{(t)}_k\right) = \mathsf{E}\left\{ -\frac{\partial^2}{\partial \hat{\phi}^{(t)2}_k} \log \left[\prod_i \mathrm{Poisson}\left(\tilde{m}^{(t)}_{uk}; \hat{\phi}^{(t)}_{uk}\right) \right] \right\} = \mathrm{diag}\left(1/\hat{\boldsymbol{\phi}}^{(t)}_k\right). \tag{5.3.17}$$

Using Eq.(5.1.1), we obtain

$$\Gamma_i(\hat{\boldsymbol{\phi}}^{(t)}_k) = \sum_j \frac{\partial}{\partial \hat{\phi}^{(t)}_{kj}} \left[\mathbf{G}\left(\hat{\boldsymbol{\phi}}^{(t)}_k\right)^{-1} \right]_{uv} = 1. \tag{5.3.18}$$

Plugging Eqs.(5.3.16;5.3.17;5.3.18) into Eq.(5.1.3) yields the SGRLD update rule as[1]

---

[1] In this paper, l is used to denote stepsize because t is used to denote time point in temporal network data.

$$\begin{aligned}
\left(\hat{\phi}_{uk}^{(t)}\right)^{*} = \Big|\hat{\phi}_{uk}^{(t)} + \epsilon_l\Big[&\left(\rho\tilde{m}_{uk}^{(t)} + \eta V\phi_{uk}^{(t-1)}\right) \\
&- \left(\rho\tilde{m}_{\cdot k}^{(t)} + \hat{\phi}_{\cdot k}^{(t)}\right)\phi_{uk}^{(t)}\Big] + \mathcal{N}(0, 2\epsilon_l\hat{\phi}_{uk}^{(t)})\Big|, \\
\{\phi_{1k}^{(t)}, \ldots, \phi_{Vk}^{(t)}\} = \{&\hat{\phi}_{1k}^{(t)}, \ldots, \hat{\phi}_{Vk}^{(t)}\}/\hat{\phi}_{\cdot k}^{(t)},
\end{aligned} \tag{5.3.19}$$

where the positiveness of $\{\hat{\phi}_{uk}^{(t)}\}_{u=1}^{V}$ is ensured by the absolute value operation $|\cdot|$. For $t = 1$, the update equation is the same except that $\eta V\phi_{uk}^{(t-1)}$ is replaced by η.

Let $\boldsymbol{\psi}_k^{(t)}$ be a nonnegative vector constrained with $\psi_{\cdot k}^{(t)} \equiv \sum_{u=1}^{V-1}\psi_{uk}^{(t)} \leq 1$. As shown in (Patterson et al. 2013), $\boldsymbol{\phi}_k^{(t)}$ can be alternatively parameterized via the reduced-mean trick as $\{\phi_{1k}^{(t)}, \ldots, \phi_{Vk}^{(t)}\} = \{\psi_{1k}^{(t)}, \ldots, \psi_{(N-1)k}^{(t)}, 1 - \psi_{\cdot k}^{(t)}\}$. Although being considered as a flawed solution because of its unstable gradients, it has been shown that this stability issue can be mitigated after preconditioning the noisy gradients (Li et al. 2016). Here, in the proposed model, we utilize the inverse of Fisher information matrix to precondition the noisy gradients, and derive an efficient update rule using the recently advanced fast sampling algorithm (Cong et al. 2017).

Given the log-posterior of $\boldsymbol{\psi}_k^{(t)}$ on the full data A as

$$\begin{aligned}
\log p(\{\psi_{uk}^{(t)}\}_{u=1}^{V-1} \mid -) \propto &\sum_{u=1}^{V-1}(\eta V\phi_{uk}^{(t-1)} + \tilde{m}_{uk}^{(t)} - 1)\log(\psi_{uk}^{(t)}) \\
&+(\eta V\phi_{Vk}^{(t-1)} + \tilde{m}_{Vk}^{(t)} - 1)\log(1 - \psi_{\cdot k}^{(t)})
\end{aligned}$$

we take the gradient of the log-posterior with respect to $\boldsymbol{\psi}_k \in \mathbb{R}_{\geq 0}^{V-1}$ on a mini-batch data scaled by $\rho \equiv |A|/|\hat{A}|$, and then we have

$$\nabla_{\boldsymbol{\psi}_k^{(t)}}[-\tilde{H}(\boldsymbol{\psi}_k^{(t)})] = \left\{\frac{\rho\tilde{m}_{uk}^{(t)} + \eta V\phi_{uk}^{(t-1)} - 1}{\psi_{uk}^{(t)}} - \frac{\rho\tilde{m}_{Vk}^{(t)} + \eta V\phi_{Vk}^{(t-1)} - 1}{1 - \psi_{\cdot k}^{(t)}}\right\}_{u=1}^{V-1}. \tag{5.3.20}$$

Note that the gradient in Eq.(5.3.20) becomes unstable if some of the components of $\boldsymbol{\psi}_k^{(t)}$ approach zeros. Nevertheless, this issue can be mitigated after preconditioning the noisy gradient with the inverse of Fisher information matrix.

Given the multinomial likelihood as

$$\{\tilde{m}_{uk}^{(t)}\}_{u=1}^{V} \sim \text{Multinomial}\left(\tilde{m}_{\cdot k}^{(t)}, \{\phi_{uk}^{(t)}\}_{u=1}^{V}\right), \tag{5.3.21}$$

we calculate the Fisher information matrix of $\boldsymbol{\psi}_k^{(t)}$ as

$$\begin{aligned}
\mathbf{G}\left(\boldsymbol{\psi}_k^{(t)}\right) &= \mathsf{E}\left\{-\frac{\partial^2}{\partial\boldsymbol{\psi}_k^{(t)2}}\log\left[\text{Multinomial}\left(\{\tilde{m}_{uk}^{(t)}\}_{u=1}^{V}\,;\,\tilde{m}_{\cdot k}^{(t)}, \{\phi_{uk}^{(t)}\}_{u=1}^{V}\right)\right]\right\} \\
&= M_k^{(t)}\left\{\text{diag}\left(\frac{1}{\boldsymbol{\psi}_k^{(t)}}\right) + \frac{\mathbf{1}\mathbf{1}^{\mathbf{T}}}{1 - \psi_{\cdot k}^{(t)}}\right\},
\end{aligned} \tag{5.3.22}$$

where $M_k^{(t)} \equiv \mathsf{E}[\tilde{m}_{\cdot k}^{(t)}]$. Using Eq.(5.1.1), we have

$$\Gamma_i(\boldsymbol{\psi}_k^{(t)}) = \sum_j \frac{\partial}{\partial\psi_{kj}^{(t)}}\left[\mathbf{G}\left(\boldsymbol{\psi}_k^{(t)}\right)^{-1}\right]_{uv} = (1 - V\psi_{uk}^{(t)})/M_k^{(t)}. \tag{5.3.23}$$

Algorithm 6 Stochastic Gradient MCMC for the proposed Dynamic Dirichlet Edge Partition Model

```
Require: temporal graphs $\{A^{(t)}\}_t$, maximum iterations $\mathcal{J}$
Ensure: posterior mean $\{\boldsymbol{\phi}_k^{(t)}\}_{k,t}, \{r_k\}_k, \{p_k\}_k, \eta$
for $l = 1{:}\mathcal{J}$ do
    Gibbs sampling on the $l$-th mini-batch for $\{\tilde{A}_{uv}^{(t)}\}_{u,v,t}$, $\{\tilde{A}_{uvk}^{(t)}\}_{u,v,k,t}$, $\{r_k\}_k, \{p_k\}_k$ and $\eta$;
    Update $\{\tilde{A}_{\cdot\cdot k}^{(t)}\}_{k,t}$, $\{\tilde{A}_{\cdot\cdot k}^{(t)}\}_{k,t}$, and $\{\tilde{A}_{\cdot\cdot k}\}_k$
    /* Update global parameters */
    for t = 1,..., T do
        Update $\{\boldsymbol{\phi}_k^{(t)}\}_k$ (Eqs. 5.3.19; 5.3.25)
    end for
end for
```

Substituting Eqs.(5.3.20;5.3.22;5.3.23) in Eq.(5.1.3), we obtain the following SGRLD update rule as

$$\left(\boldsymbol{\psi}_k^{(t)}\right)^* = \Bigg\{\boldsymbol{\psi}_k^{(t)} + \frac{\epsilon_l}{M_k^{(t)}}\Big[\left(\rho\tilde{\mathbf{m}}_k^{(t)} + \eta V\tilde{\boldsymbol{\phi}}_k^{(t-1)}\right) - \left(\tilde{m}_{\cdot k}^{(t)} + \eta V\right)\boldsymbol{\psi}_k^{(t)}\Big] + \mathcal{N}\left(0, \frac{2\epsilon_l}{M_k^{(t)}}\left[\mathrm{diag}(\boldsymbol{\psi}_k^{(t)}) - \boldsymbol{\psi}_k^{(t)}\boldsymbol{\psi}_k^{(t)\mathrm{T}}\right]\right)\Bigg\}_{\angle}, \tag{5.3.24}$$

where $\tilde{\boldsymbol{\phi}}_k^{(t)} \equiv [\phi_{1k}^{(t)}, \ldots, \phi_{(N-1)k}^{(t)}]$, and $\{\cdot\}_{\angle}$ denotes the constraint that $\psi_{uk}^{(t)} \geq 0$, and $\sum_{u=1}^{V-1}\psi_{uk}^{(t)} \leq 1$.

It is computational expensive to simulate the multivariate normal distribution in Eq.(5.3.24) using Cholesky decomposition. Therefore, we resort to a recently advanced fast sampling algorithm (Cong et al. 2017). Instead of updating $\boldsymbol{\psi}_k^{(t)}$, we can equivalently update $\boldsymbol{\phi}_k^{(t)}$ that is drawn from a related multivariate normal distribution with a diagonal covariance matrix as

$$\left(\boldsymbol{\phi}_k^{(t)}\right)^* = \Bigg\{\boldsymbol{\phi}_k^{(t)} + \frac{\epsilon_l}{M_k^{(t)}}\Big[\left(\rho\tilde{\mathbf{m}}_k^{(t)} + \eta V\tilde{\boldsymbol{\phi}}_k^{(t-1)}\right) - \left(\tilde{m}_{\cdot k}^{(t)} + \eta V\right)\boldsymbol{\phi}_k^{(t)}\Big] + \mathcal{N}\left(0, \frac{2\epsilon_l}{M_k^{(t)}}\left[\mathrm{diag}(\boldsymbol{\phi}_k^{(t)})\right]\right)\Bigg\}_{\angle}. \tag{5.3.25}$$

For $t = 1$, we replace $\eta V\tilde{\boldsymbol{\phi}}_k^{(t-1)}$ by η in the update rule. Our SG-MCMC algorithm iteratively updates the parameters $\{\boldsymbol{\phi}_k^{(t)}\}_{k,t}$ and samples the remaining ones as in the proposed Gibbs sampler.

The main procedure is summarized in Algorithm 6.

5.4 EXPERIMENTS

We now present the experimental results on several real-world datasets to evaluate the accuracy and efficiency of the proposed model. The proposed model is referred to as D^{2}EPM (the **D**irichlet **D**ynamic **E**dge **P**artition **M**odel) with Gibbs sampler, Expanded-Mean SGRLD and Reduced-Mean SGRLD, as D^{2}EPM-Gibbs, D^{2}EPM-EM-SGRLD, D^{2}EPM-RM-SGRLD, respectively. We compare our model with two baselines: (1) the dynamic stochastic block model (DSBM) (K. S. Xu et al. 2014). (2) the gamma process dynamic network model (GaP-DNM) that captures the evolution of nodes' memberships using a gamma Markov chain construction (S. Yang and H. Koeppl

Method	Hypertext	Facebook Like	Facebook Message	NIPS
DSBM	0.703	0.848	0.814	0.899
GaP-DNM	0.766	0.887	0.888	0.887
D^2EPM-Gibbs	**0.812**	**0.912**	**0.929**	0.895
D^2EPM-EM-SGRLD	0.808	0.871	0.926	0.902
D^2EPM-RM-SGRLD	0.809	0.868	0.927	**0.916**

Table 5.1: Link prediction on temporal network data. We report the averaged area under the ROC curve (AUROC) over five different training/test partitions, and highlight the best scores in bold.

2018b). We also compare the proposed SG-MCMC algorithms in terms of link prediction accuracy vs wall-clock run time.

We chose the following datasets in our experiments:

1. **Hypertext:** This dataset (Mastrandrea et al. 2015) contains the interactions between 113 participants at the 2009 Hypertext conference. We generated a dynamic network assuming each hour as a snapshot, and creating an edge between each pair of participants at time t if they have at least one contact recorded during that snapshot.
2. **Facebook Like[2]:** This dataset contains 33,720 broadcast messages among 899 students over 7 months from a Facebook-like forum. We generated a dynamic network aggregating the data into monthly snapshots, and creating an edge between each pair of nodes if the presence of messages between them is recorded during that snapshot.
3. **Facebook Message[3]:** This dataset contains 59,835 private messages among 1,899 college students over 7 months. We generated a dynamic network aggregating the data into monthly snapshots, and creating an edge between each pair of nodes if the presence of messages between them is recorded during that snapshot.
4. **NIPS Co-authorship[4]:** This dataset contains 4,798 publications by 5,722 authors in the NIPS conference over 10 years. We generated a dynamic network aggregating the data into yearly snapshots and creating an edge between two authors in a snapshot if they appear on the same publication in that year.

First, we compared the accuracy of all the models in terms of link prediction. We trained all the methods using 80% of randomly chosen entries (either links or non-links) in the given network data, and used the remaining 20% as the held-out data to test the trained model. Each experiment is conduced five times with different training/test partitions, and the averaged Area Under the Receiver Operating Characteristi curve (AUROC) for all the data sets is reported in the final results. The proposed method is implemented in MATLAB. Unless specified otherwise, we initialized the GaP-DNM and the proposed D^2EPM with $K = 50$ because both two models can automatically determine the number of communities. We ran both GaP-DNM and D^2EPM-Gibbs for 3000 iterations with 2000 burn-in and 1000 collection iterations. We set the hyperparameters as $g_k = 0.1, a_0 = b_0 = 0.01, c_0 = 1$ and $\alpha = 1/K$. A sensitivity analysis revealed that we obtain similar results when instead setting $g_k = 0.01$ or 1. The SG-MCMC algorithms were also run for

---

[2] `https://tinyurl.com/ycdezko6`.

[3] `https://snap.stanford.edu/data/CollegeMsg.html`.

[4] `http://www.cs.huji.ac.il/~papushado/nips_collab_data.html`.

Method	Hypertext	Facebook Like	Facebook Message	NIPS
GaP-DNM	0.751	0.953	2.436	12.759
D^{2}EPM-Gibbs	0.745	0.755	1.597	9.383
D^{2}EPM-EM-SGRLD	0.317	0.580	0.969	4.931
D^{2}EPM-RM-SGRLD	0.395	0.485	1.032	4.351

Table 5.2: Comparison of computation time (seconds).

the same number of iterations, with mini-batch size equal to one-fourth of the number of nonzero edges in the training data. We used the stepsize $\epsilon_l = (a(1 + l/b))^{-c}$ and the optimal parameters a, b, c as in (Patterson et al. 2013; Ma et al. 2015). All the experiments are conducted on a standard computer with 24 GB RAM.

Table 5.1 shows the experimental results on the link prediction task. Overall, the D^{2}EPM (both Gibbs sampling and SG-MCMC algorithms) outperform the other baselines. Specifically, the sampling based methods (GaP-DNM and D^{2}EPM-Gibbs) achieve better accuracy than the DSBM based on the extended Kalman filter on the relatively small datasets although the former two methods require a sufficiently large number of iterations to converge. For the medium-sized NIPS dataset, the SG-MCMC algorithms perform better than the Gibbs sampling, suggesting the batch Gibbs sampler mixes poorly. Using the extended Kalman filter to perform inference, DSBM is much faster than the probabilistic models. We report per-iteration computation time of GaP-DNM, D^{2}EPM-Gibbs, D^{2}EPM-EM-SGRLD and D^{2}EPM-RM-SGRLD with Matlab/MEX/C implementation on all these datasets in Table. 5.2. In Figure 5.1, we compare the AUROC vs wall-clock run time for D^{2}EPM-Gibbs, D^{2}EPM-EM-SGRLD, D^{2}EPM-RM-SGRLD on Facebook message and NIPS datasets. For Facebook message dataset, we found that batch Gibbs sampler converges very fast, and the SG-MCMC algorithms converge to the same level of accuracy in comparable time. For the larger network (NIPS dataset), the SG-MCMC algorithms converge much faster than the Gibbs sampler.

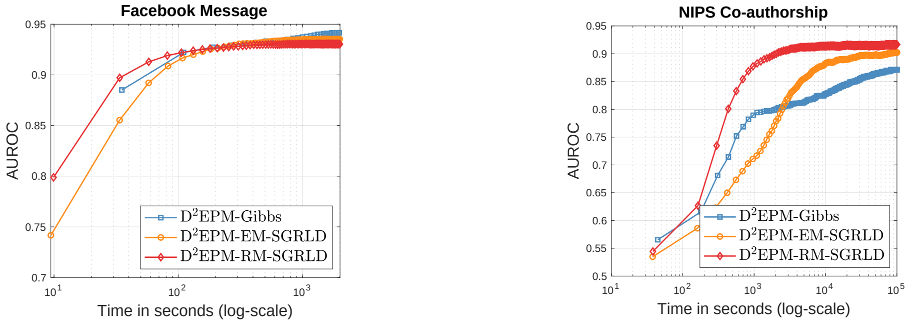


Figure 5.1: Running time comparison of D^{2}EPM-Gibbs, D^{2}EPM-EM-SGRLD, D^{2}EPM-RM-SGRLD on Facebook message (left) and NIPS Co-authorship datasets (right).

5.5 CONCLUSIONS

We presented a novel dynamic edge partition model for temporal relational learning by capturing the evolution of nodes' memberships over time using a Dirichlet Markov chain construction.

The appropriate number of latent communities is automatically inferred via the hierarchical beta-gamma prior. In particular, the new framework admits a simple-to-implement Gibbs sampling scheme using the negative-binomial augmentation technique, and also enables us to develop a scalable inference algorithm based on the SG-MCMC framework. We demonstrate the accuracy and efficiency of the novel methods on several real-world datasets. The proposed framework allows us to incorporate available node-specific side information via the Pólya-Gamma augmentation technique (Polson et al. 2013), and also to infer a tree-structured latent communities hierarchy using the gamma belief-net (M. Zhou et al. 2016).

6

NONPARAMETRIC BAYESIAN GROUP FACTOR ANALYSIS

Factor analysis (FA) is a powerful tool widely used to infer low-dimensional structure in multivariate data. More specifically, FA models attempt to represent a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ by the product of two matrices plus residual noise as

$$\mathbf{X} = \mathbf{F}\mathbf{G} + \mathbf{E},$$

where $\mathbf{F} \in \mathbb{R}^{N \times K}$ denotes the factor score matrix, and $\mathbf{G} \in \mathbb{R}^{K \times D}$ denotes the factor loading matrix; $\mathbf{E} \in \mathbb{R}^{N \times D}$ is the residual noise matrix. For high-dimensional data, FA models imposing sparsity-inducing priors (West 2003; Rai and Daume III 2008; Paisley and L. Carin 2009; Knowles and Z Ghahramani 2011) or regularizations (Zou, Hastie, and Tibshirani 2006; Witten, Hastie, and Tibshirani 2009) over the inferred loading matrices are developed to improve interpretability of the inferred low-dimensional structure. For example, in gene expression analysis, a factor loading matrix characterizing the connections between transcription factors and regulated genes are expected to be sparse (Carvalho et al. 2008).

In many real-world applications, we often deal with multiple related datasets – each comprising a group of variables – that need to be factorized in a common subspace. For instance, latent Dirichlet allocation (D. M. Blei et al. 2003) and Poisson factor analysis models (M. Zhou and L. Carin 2015b) have been developed to learn the shared latent topics among multiple documents. Recently, group factor analysis (GFA) models (Virtanen et al. 2012; Bunte et al. 2016) using the automatic relevance determination (ARD) prior have been proposed for drug sensitivity prediction and functional neuroimaging. However, the modeling flexibility achieved by these GFA models comes at a price as their inference usually requires Markov chain Monte Carlo (MCMC) to perform posterior computation, which makes them to scale poorly for large-scale GFA problems. Alternatively, variational Bayesian inference has been shown to be efficient for large-scale data analysis by making an independence assumption among latent variables and parameters (Wainwright and Jordan 2008). However, this strong assumption may lead to very inaccurate results in practical applications, especially for GFA problems where latent variables might be tightly coupled.

Motivated by this limitation, we propose a computationally efficient collapsed variational inference algorithm for the nonparametric Bayesian group factor analysis (NGFA) model. The proposed NGFA model is built upon the hierarchical beta process (HBP) (Thibaux et al. 2007). We note that the HBP has been investigated in (B. Chen et al. 2011; Gupta et al. 2012a; Gupta et al. 2012b) for joint modeling of multiple data matrices utilizing MCMC, but again showed poor scalability and slow convergence. For nonparametric Bayesian models, such as the hierarchical Dirichlet process (HDP) topic model (Teh et al. 2007) and the hierarchical Dirichlet process hidden Markov model (Fox et al. 2011), collapsed Gibbs sampling (CGS) algorithms are typically employed to perform posterior computation because CGS rapidly convergences onto the true posterior. However, it remains challenging to assess the convergence of CGS algorithms for practical use. To address this issue, collapsed variational inference algorithms (Teh et al. 2006; Teh et al. 2008; J. Foulds et al. 2013) are developed for topic models by integrating out model parameters, and then applying the mean field approximation to the latent variables. Recently, collapsed variational inference algorithms have been developed for hidden Markov models (Wang et al. 2013), nonparametric relational models (Ishiguro et al. 2017) and Markov jump processes (B. Zhang et al. 2017) with encouraging results.

In this chapter, we aim to develop a nonparametric Bayesian group factor analysis (NGFA) model, and a collapsed variational inference algorithm to perform fast inference for the developed NGFA. We make the following contributions:

- We tackle the group factor analysis problems using a Bayesian nonparametric method based on the hierarchical beta Bernoulli process. The total number of factors is automatically learned from data. Specifically, the NGFA model induces both group-wise and element-wise structured sparsity effectively compared to state-of-the-art GFA methods.
- An efficient collapsed variational inference algorithm is proposed to infer the NGFA model.
- We apply the developed method to real world multiple related datasets, with encouraging results.

This chapter is organized as follows. In Section 5.1, we describe the nonparametric Bayesian group factor analysis model. The developed collapsed variational inference algorithm for the NGFA is introduced in Section 5.2. Experimental results are presented in Section 5.3. Finally, conclusions and possible directions for future research are discussed in Section 5.4.

6.1 NONPARAMETRIC BAYESIAN GROUP FACTOR ANALYSIS

Given multiple related data matrices $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}$, each with N samples, i.e., $\mathbf{X}^{(m)} \in \mathbb{R}^{N \times D_m}$, our goal is to factorize each dataset $\mathbf{X}^{(m)}$ into the product of a common factor matrix $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K]$ of size $N \times K$, and a group-specific factor loading matrix $\mathbf{G}^{(m)}$ of size $K \times D_m$ as

$$\mathbf{X}^{(m)} = \mathbf{F}\mathbf{G}^{(m)} + \mathbf{E}^{(m)}, \tag{6.1.1}$$

where $\mathbf{E}^{(m)} = [\mathbf{e}_1^{(m)}, \ldots, \mathbf{e}_{D_m}^{(m)}]$ is assumed to be Gaussian noise for the m-th dataset or group. We impose independent normal priors over $\mathbf{e}_d^{(m)} \in \mathbb{R}^N$, i.e., $\mathbf{e}_d^{(m)} \sim \mathcal{N}(0, \operatorname{diag}(\tau_1^{(m)}, \ldots, \tau_N^{(m)}))$, where $\tau_n^{(m)}$ controls the variance of N-th sample in the m-th group. As commonly used in factor analysis (Rai and Daume III 2008; Paisley and L. Carin 2009; Knowles and Z Ghahramani 2011), we put a normal prior on each factor $\mathbf{f}_k$, i.e., $\mathbf{f}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, where $\mathbf{I}_N$ is an identity matrix of size N. To explicitly capture the sparsity, we model the factor loading matrix $\mathbf{G}^{(m)}$ for each group by the element-wise product of a binary matrix $\mathbf{Z}^{(m)}$ and a real-valued weight matrix $\mathbf{W}^{(m)}$, i.e., $\mathbf{G}^{(m)} = \mathbf{Z}^{(m)} \odot \mathbf{W}^{(m)}$. More specifically, we place a normal prior over each element of $\mathbf{W}^{(m)}$, i.e., $w_{kd}^{(m)} \sim \mathcal{N}(0, (\lambda_{kd}^{(m)})^{-1})$. To allow the number of factors K to be automatically inferred from data, we model each row of $\mathbf{Z}^{(m)}$ as a draw from a group-specific Bernoulli process.

As our goal is to factorize multiple related data matrices using a common set of factors, we naturally consider the hierarchical beta process (Thibaux et al. 2007) that allows us to generate a set of latent factors from a *global* beta process B, and then allow the generated factors to be shared among all the groups. The usage of the generated factors in each group is determined by the group-specific beta process $A^{(m)}$. More specifically, the (truncated) hierarchical beta Bernoulli process is

$$\begin{aligned}
\mathbf{f}_k &\sim \mathcal{N}(0, \mathbf{I}_N), \qquad (6.1.2)\\
\beta_k &\sim \text{Beta}(\kappa_0/K, \kappa_0(K-1)/K),\\
\pi_k^{(m)} &\sim \text{Beta}(\alpha^{(m)}\beta_k, \alpha^{(m)}\bar{\beta}_k),\\
B &\equiv \sum_{k=1}^{K} \beta_k \delta_{\mathbf{f}_k},\\
A^{(m)} &\equiv \sum_{k=1}^{K} \pi_k^{(m)} \delta_{\mathbf{f}_k},\\
z_{kd}^{(m)} &\sim \text{Bernoulli}(\pi_k^{(m)}),
\end{aligned}$$

where $\bar{\beta}_k \equiv 1 - \beta_k$, and K is a truncation level that is set sufficiently large to ensure a good approximation to the truly infinite model. The concentration parameters of the global beta process and the local group-specific beta process are κ_0 and $\alpha^{(m)}$, respectively. The total number of factors shared among all groups is determined by κ_0, and the amount of variability of each $A^{(m)}$ around B is determined by $\alpha^{(m)}$. To improve the flexility of the model, we place gamma priors on $\lambda_{kd}^{(m)}$, $\tau_n^{(m)}$ and $\alpha^{(m)}$, respectively, as $\lambda_{kd}^{(m)} \sim \text{Gamma}(g_0, h_0)$, $\tau_n^{(m)} \sim \text{Gamma}(e_0, f_0)$, $\alpha^{(m)} \sim \text{Gamma}(c_0, d_0)$.

The graphical representation of the NGFA model is shown in shown in Fig. 6.1 (top).

6.2 COLLAPSED VARIATIONAL INFERENCE

The main idea of collapsed variational inference is to marginalize out model parameters, and then apply the mean field method to approximate the distribution over latent variables. We note that marginalizing out the parameters induces dependencies among the latent variables. However, each latent variable interacts with the remaining variables only through the sufficient statistics (i.e. the field) in the collapsed space, and the influence of any single variable on the field is small. Hence, the dependency between any two latent variables is weak, suggesting that the mean field assumption is better justified in the collapsed space. In our case, we first marginalize out the group-specific beta process parameters to obtain the marginal distribution over latent variables. We then employ the variational posterior to approximate the distribution of latent variables and the remaining parameters.

Notation. When expressing the conditional distribution, we will use the shorthand "–" to denote full conditionals, i.e., all other variables. For the sake of clarity, we use $\mathbf{X}$ to denote the set of matrices $\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}\right)$. Similarly, let $\mathbf{Z}$ denote $\left(\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}\right)$, and $\boldsymbol{\pi}$ denote $\left(\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(M)}\right)$. We repeatedly exploit the following three results (Teh et al. 2008) to derive the collapsed variational inference algorithm for the NGFA.

Result 1. The geometric expectation of a non-negative random variable y is defined as $\mathsf{G}[y] \equiv \exp(\mathsf{E}[\ln(y)])$. If y is gamma distributed, i.e., $f_Y(y \mid a, b) \propto y^{a-1}e^{-by}$, the geometric expectation of y is $\mathsf{G}[y] = \frac{\exp(\Psi(a))}{b}$, where $\Psi(y) = \frac{\partial \ln \Gamma(y)}{\partial y}$ is the digamma function. For a beta distributed random variable y, i.e., $f_Y(y \mid a, b) \propto y^{a-1}(1-y)^{b-1}$, the geometric expectation of y is $\mathsf{G}[y] = \frac{\exp[\Psi(a)]}{\exp[\Psi(a+b)]}$. If $y_1, \ldots, y_K$ are mutually independent, we have, $\mathsf{G}\left[\prod_{k=1}^{K} y_k\right] = \prod_{k=1}^{K} \mathsf{G}[y_k]$.

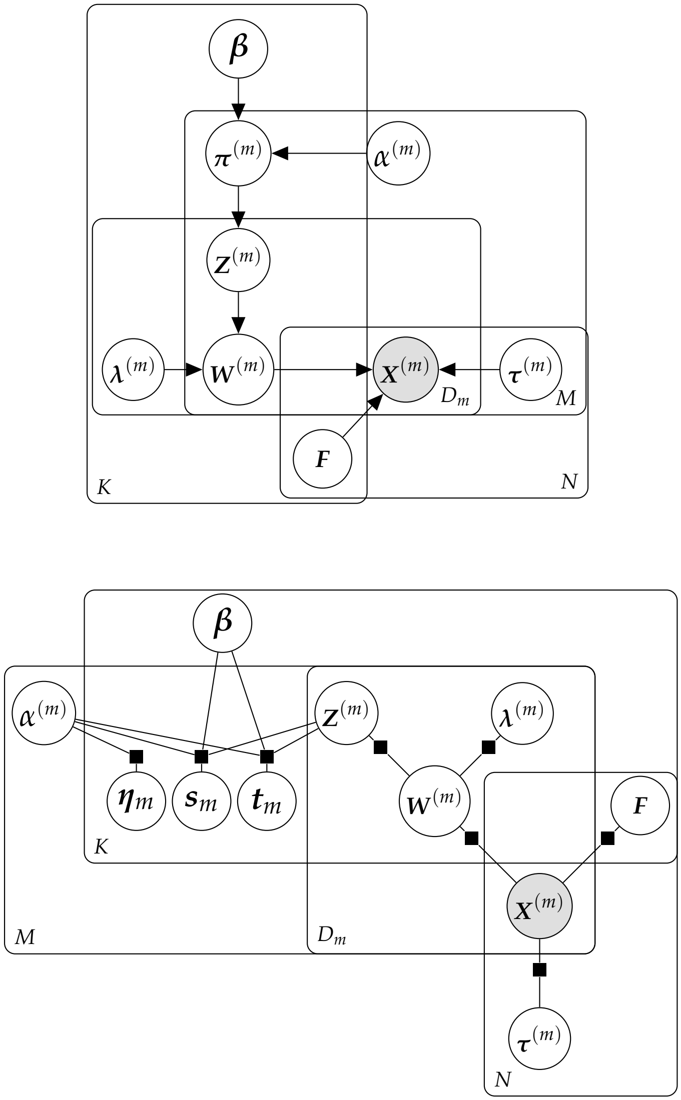


Figure 6.1: Top: The graphical representation of the proposed model. Bottom: Factor graph of the model with auxiliary variables.

Result 2. According to the central limit theorem, if y is the sum of N independent Bernoulli random variables, i.e., $y = \sum_{i=1}^{N} u_i$, where $u_i \sim \text{Bernoulli}(\xi_i)$, then for large enough N, y is well approximated by a Gaussian random variable with mean and variance as

$$\mathsf{E}[y] = \sum_{i=1}^{N} \xi_i, \qquad \mathsf{V}[y] = \sum_{i=1}^{N} \xi_i (1 - \xi_i),$$

respectively. Moreover, the expectation of $\ln(y)$ can be approximated using the second-order Taylor expansion (Hoef 2012) as

$$\mathsf{E}[\ln(y)] \approx \ln(\mathsf{E}[y]) - \frac{\mathsf{V}[y]}{2(\mathsf{E}[y])^2}.$$

Result 3. If l is the sum of independent Bernoulli random variables, i.e., $l = \sum_i u_i$, where $u_i \sim \text{Bernoulli}(\xi_i)$, we use $p_+(l)$ to denote the probability of l being positive, i.e.,

$$\begin{aligned} p_+(l) \equiv p(l > 0) &= 1 - \prod_i p(u_i = 0) \\ &= 1 - \exp\left[\sum_i \ln(1 - \xi_i)\right]. \end{aligned}$$

Accordingly, the expectation and variance conditional on $l > 0$ are defined as $\mathsf{E}_+[l] \equiv \frac{\mathsf{E}[l]}{p_+(l)}$ and $\mathsf{V}_+[l] \equiv \frac{\mathsf{V}[l]}{p_+(l)}$, respectively. If y is then a Chinese restaurant table (CRT) (Pitman 2006) distributed random variable, i.e., $f_Y(y \mid a, l) = \frac{\Gamma(a)}{\Gamma(a+l)} \left[{l \atop y}\right] a^y$, where $y = 0, 1, \ldots, l$, and $\left[{n \atop m}\right]$ denoting the unsigned Stirling number of the first kind, then the expectation of y can be closely approximated using the improved second-order Taylor expansion as

$$\begin{aligned} \mathsf{E}[y] \approx \mathsf{G}[a] p_+(l) \Big(& \Psi(\mathsf{G}[a] + \mathsf{E}_+[l]) \\ & - \Psi(\mathsf{G}[a]) + \frac{\mathsf{V}_+[l] \Psi'(\mathsf{G}[a] + \mathsf{E}_+[l])}{2} \Big), \end{aligned}$$

where $\Psi'(y) = \frac{\partial^2 \ln \Gamma(y)}{\partial y^2}$ is the trigamma function.

6.2.1 *Collapsed representation*

First, we describe how to obtain the marginal distribution of latent variables. In the next subsection, we will then describe how to derive the CVI algorithm in the collapsed space.

For the NGFA introduced in the previous section, integrating out $\boldsymbol{\pi}$ using the beta-Bernoulli conjugacy yields the marginal distribution of $\mathbf{Z}$ as

$$p(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = \int p(\mathbf{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\pi} = \prod_{m,k} \frac{\Gamma(\alpha^{(m)})}{\Gamma(\alpha^{(m)} + D_m)} \frac{\Gamma(\alpha^{(m)} \beta_k + \hat{n}_{mk})}{\Gamma(\alpha^{(m)} \beta_k)} \frac{\Gamma(\alpha^{(m)} \bar{\beta}_k + \tilde{n}_{mk})}{\Gamma(\alpha^{(m)} \bar{\beta}_k)} \tag{6.2.1}$$

where we define $\hat{n}_{mk} \equiv \sum_d \mathbb{1}(z_{kd}^{(m)} = 1)$ and $\tilde{n}_{mk} \equiv \sum_d \mathbb{1}(z_{kd}^{(m)} = 0)$, and $\mathbb{1}(\cdot)$ is the standard indicator function.

As the ratios of gamma functions in Eq. 6.2.1 give rise to difficulties for updating hyperparameter posteriors, we augment the marginal distribution **Z** by introducing three sets of auxiliary variables. More specifically, using the auxiliary variable method (Teh et al. 2007), the first ratio of gamma function can be re-expressed as

$$\frac{\Gamma(\alpha^{(m)})}{\Gamma(\alpha^{(m)}+D_m)} = \frac{1}{\Gamma(D_m)} \int_0^1 \eta_m^{\alpha^{(m)}} (1-\eta_m)^{D_m-1} \left(1+\frac{D_m}{\alpha^{(m)}}\right) \mathrm{d}\eta_m. \tag{6.2.2}$$

Via the relation between the gamma function and the Stirling numbers of the first kind (Teh et al. 2007), the second and third ratio of gamma functions can be re-expressed, respectively, as

$$\frac{\Gamma(\alpha^{(m)}\beta_k+\hat{n}_{mk})}{\Gamma(\alpha^{(m)}\beta_k)} = \sum_{s_{mk}=0}^{\hat{n}_{mk}} \begin{bmatrix}\hat{n}_{mk}\\ s_{mk}\end{bmatrix} (\alpha^{(m)}\beta_k)^{s_{mk}}, \tag{6.2.3}$$

$$\frac{\Gamma(\alpha^{(m)}\bar{\beta}_k+\tilde{n}_{mk})}{\Gamma(\alpha^{(m)})} = \sum_{t_{mk}=0}^{\tilde{n}_{mk}} \begin{bmatrix}\tilde{n}_{mk}\\ t_{mk}\end{bmatrix} (\alpha^{(m)}\bar{\beta}_k)^{t_{mk}}. \tag{6.2.4}$$

Substituting (Eqs. 6.2.2; 6.2.3; 6.2.4) into Eq. 6.2.1, we immediately obtain the joint distribution of the latent and auxiliary variables as

$$\begin{aligned} p(\mathbf{Z},\mathbf{s},\mathbf{t},\boldsymbol{\eta} \mid \boldsymbol{\beta},\boldsymbol{\alpha}) \propto \prod_{m,k} & \eta_m^{\alpha^{(m)}-1}(1-\eta_m)^{D_m-1} \\ & \times \begin{bmatrix}\hat{n}_{mk}\\ s_{mk}\end{bmatrix} (\alpha^{(m)}\beta_k)^{s_{mk}} \begin{bmatrix}\tilde{n}_{mk}\\ t_{mk}\end{bmatrix} (\alpha^{(m)}\bar{\beta}_k)^{t_{mk}}. \end{aligned} \tag{6.2.5}$$

The factor graph of the expanded system with auxiliary variables is shown in Fig. 6.1 (bottom). The conditional distribution of a single latent variable $z_{kd}^{(m)}$ can be derived using the marginal distribution of **Z** and the likelihood function according to Eq. 6.1.1 as

$$\begin{aligned} p(z_{kd}^{(m)}=1 \mid -) \propto & \exp\left[\ln(\alpha^{(m)}\beta_k+\hat{n}_{km}^{\neg d})\right] \\ & \times \exp\left[-\frac{1}{2}\sum_n \tau_n^{(m)} \left(\left(w_{kd}^{(m)}\right)^2 f_{nk}^2 - 2w_{kd}^{(m)}\ \tilde{x}_{nd}^{(m)\ \neg k}\right)\right], \end{aligned} \tag{6.2.6}$$

where $\left(\tilde{x}_{nd}^{(m)}\right)^{\neg k} \equiv \left(x_{nd}^{(m)} - \sum_{j\neq k} z_{jd}^{(m)} w_{jd}^{(m)} f_{nj}\right)$, and $\hat{n}_{km}^{\neg d} \equiv \sum_{d'\neq d} \mathbb{1}(z_{kd'}^{(m)}=1)$.

6.2.2 *Variational approximation*

Next, we shall introduce the variational approximation for our expanded system. For the sake of simplicity, the remaining parameters $(\mathbf{W},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\tau},\boldsymbol{\alpha})$ are denoted by $\boldsymbol{\theta}$. Formally, the variational posterior over the augmented variables system is assumed to be of the form

$$q(\mathbf{Z},\boldsymbol{\theta},\mathbf{s},\mathbf{t},\boldsymbol{\eta}) = q(\boldsymbol{\theta})q(\mathbf{s},\mathbf{t},\boldsymbol{\eta} \mid \mathbf{Z})q(\mathbf{Z}),$$

where $q(\boldsymbol{\theta}) \equiv q(\mathbf{W})q(\mathbf{F})q(\boldsymbol{\beta})q(\boldsymbol{\lambda})q(\boldsymbol{\tau})q(\boldsymbol{\alpha})$, and we define the variational posterior for each parameter as

$$
\begin{aligned}
q(\mathbf{W}) &= \prod_{m,d,k} \mathcal{N}(w_{kd}^{(m)}; \mu_{w_{kd}}^{(m)}, \sigma_{w_{kd}}^{(m)}),\\
q(\mathbf{F}) &= \prod_{n,k} \mathcal{N}(f_{nk}; \mu_{f_{nk}}, \sigma_{f_{nk}}),\\
q(\boldsymbol{\beta}) &= \prod_{k} \mathrm{Beta}(\beta_k; a_k, b_k),\\
q(\boldsymbol{\lambda}) &= \prod_{m,d,k} \mathrm{Gamma}(\lambda_{kd}^{(m)}; e_{kd}^{(m)}, f_{kd}^{(m)}),\\
q(\boldsymbol{\tau}) &= \prod_{m,n} \mathrm{Gamma}(\tau_n^{(m)}; g_n^{(m)}, h_n^{(m)}),\\
q(\boldsymbol{\alpha}) &= \prod_{m} \mathrm{Gamma}(\alpha^{(m)}; c^{(m)}, d^{(m)}),\\
q(\mathbf{Z}) &= \prod_{m,d,k} \mathrm{Bernoulli}(z_{kd}^{(m)}; \rho_{kd}^{(m)}),\\
q(\mathbf{s}|\mathbf{Z}) &= \prod_{m,k} \begin{bmatrix} \hat{n}_{mk} \\ s_{mk} \end{bmatrix} (\mathsf{G}[\alpha^{(m)}\beta_k])^{s_{mk}},\\
q(\mathbf{t}|\mathbf{Z}) &= \prod_{m,k} \begin{bmatrix} \tilde{n}_{mk} \\ t_{mk} \end{bmatrix} (\mathsf{G}[\alpha^{(m)}(1-\beta_k)])^{t_{mk}},\\
q(\boldsymbol{\eta}|\mathbf{Z}) &= \prod_{m} \mathrm{Beta}(\eta_m; \mathsf{E}\left[\alpha^{(m)}\right], D_m).
\end{aligned}
$$

Note that the true posterior $p(\mathbf{s}, \mathbf{t}, \boldsymbol{\eta} \mid \mathbf{Z})$ is used in our variational update subsequently.

Evidence Lower Bound (ELBO): The log marginal likelihood of data is lower bounded as

$$
\begin{aligned}
\log p(\mathbf{X} \mid \kappa_0) &\geq \mathsf{E}\left[p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{s}, \mathbf{t}, \boldsymbol{\eta} \mid \kappa_0)\right] - \mathsf{E}\left[q(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{s}, \mathbf{t}, \boldsymbol{\eta})\right]\\
&= \mathsf{E}_{q(\boldsymbol{\theta},\mathbf{Z})}\left[\mathsf{E}_{q(\mathbf{s},\mathbf{t},\boldsymbol{\eta}|\mathbf{Z})}\left[\log \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{s}, \mathbf{t}, \boldsymbol{\eta} \mid \kappa_0)}{q(\mathbf{s}, \mathbf{t}, \boldsymbol{\eta} \mid \mathbf{Z})}\right] - \log q(\boldsymbol{\theta}, \mathbf{Z})\right]\\
&= \mathsf{E}_{q(\boldsymbol{\theta},\mathbf{Z})}\left[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta} \mid \kappa_0) - q(\boldsymbol{\theta}, \mathbf{Z})\right], \qquad (6.2.7)
\end{aligned}
$$

where the second equality holds provided that $q(\mathbf{s}, \mathbf{t}, \boldsymbol{\eta} \mid \mathbf{Z})$ is set to its true posterior. To derive the variational update for each parameter, we expand the ELBO for each term in Eq. 6.2.7 as

$$
\begin{aligned}
&\log p(\mathbf{X} \mid \kappa_0) \geq \ \mathsf{E}\left[\log p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \mathbf{F}, \boldsymbol{\tau})\right]\\
&+ \mathsf{E}\left[\log p(\mathbf{W})\right] - \mathsf{E}\left[\log q(\mathbf{W})\right] + \mathsf{E}\left[\log p(\mathbf{Z})\right] - \mathsf{E}\left[\log q(\mathbf{Z})\right]\\
&+ \mathsf{E}\left[\log p(\mathbf{F})\right] - \mathsf{E}\left[\log q(\mathbf{F})\right] + \mathsf{E}\left[\log p(\boldsymbol{\lambda})\right] - \mathsf{E}\left[\log q(\boldsymbol{\lambda})\right]\\
&+ \mathsf{E}\left[\log p(\boldsymbol{\tau})\right] - \mathsf{E}\left[\log q(\boldsymbol{\tau})\right] + \mathsf{E}\left[\log p(\boldsymbol{\alpha})\right] - \mathsf{E}\left[\log q(\boldsymbol{\alpha})\right]\\
&+ \mathsf{E}\left[\log p(\boldsymbol{\beta})\right] - \mathsf{E}\left[\log q(\boldsymbol{\beta})\right]. \qquad (6.2.8)
\end{aligned}
$$

The variational updates for each parameter are obtained by taking the derivate of the ELBO in Eq. 6.2.8 w.r.t. each parameter and setting it to zero.

Updating $q(\mathbf{Z})$: The variational update for each latent variable $z_{kd}^{(m)}$ is

$$\begin{aligned} q(z_{kd}^{(m)} = 1) &\propto \exp\left(\mathsf{E}_{q(\mathbf{Z},\boldsymbol{\theta} \setminus z_{kd}^{(m)})}\left[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta} \mid \kappa_0)\right]\right) \\ &\propto \exp\left(\mathsf{E}_{q(\mathbf{Z},\boldsymbol{\theta} \setminus z_{kd}^{(m)})}\left[\ln p(z_{kd}^{(m)} = 1 \mid -)\right]\right), \end{aligned} \tag{6.2.9}$$

where $(\mathbf{Z}, \boldsymbol{\theta} \setminus z_{kd}^{(m)})$ means all the variables and parameters excluding $z_{kd}^{(m)}$.

Plugging Eq. 6.2.6 into Eq. 6.2.9, we obtain the variational update for $q(z_{kd}^{(m)} = 1)$ as

$$\begin{aligned} &q(z_{kd}^{(m)} = 1) \qquad (6.2.10) \\ &\propto \exp\left\{\mathsf{E}\left[\ln\left(\alpha^{(m)}\beta_k + \hat{n}_{mk}^{\neg d}\right)\right] - \frac{1}{2}\sum_n \mathsf{E}[\tau_n^{(m)}]\left(\mathsf{E}[(w_{kd}^{(m)})^2]\mathsf{E}[f_{nk}^2] - 2\mathsf{E}[w_{kd}^{(m)}]\mathsf{E}[f_{nk}]\,\tilde{x}_{nd}^{(m)\,\neg k}\right)\right\}. \end{aligned}$$

The exact computation of the log count in Eq. 6.2.10 is too expensive in practice. According to Result 2, we can approximate it as

$$\begin{aligned} \mathsf{E}\left[\ln\left(\alpha^{(m)}\beta_k + \hat{n}_{mk}^{\neg d}\right)\right] \approx{}& \ln\left(\mathsf{G}[\alpha^{(m)}\beta_k] + \mathsf{E}\left[\hat{n}_{mk}^{\neg d}\right]\right) \\ &- \frac{\mathsf{V}\left[\hat{n}_{mk}^{\neg d}\right]}{2\left(\mathsf{G}[\alpha^{(m)}\beta_k] + \mathsf{E}\left[\hat{n}_{mk}^{\neg d}\right]\right)^2}, \end{aligned}$$

where the mean and variance of $\hat{n}_{mk}^{\neg d}$ are given by

$$\begin{aligned} \mathsf{E}\left[\hat{n}_{mk}^{\neg d}\right] &= \sum_{d' \neq d} q(z_{kd}^{(m)} = 1), \\ \mathsf{V}\left[\hat{n}_{mk}^{\neg d}\right] &= \sum_{d' \neq d} q(z_{kd}^{(m)} = 1) q(z_{kd}^{(m)} = 0). \end{aligned}$$

Updating auxiliary variables: Now we explain how to update the auxiliary variables efficiently using Gaussian approximation techniques. The variational posteriors for the auxiliary variables $\boldsymbol{\eta}$ is

$$q(\boldsymbol{\eta} \mid \mathbf{Z}) \propto \prod_m \eta_m^{\mathsf{E}[\alpha^{(m)}]-1}(1-\eta_m)^{D_m-1}.$$

As $\boldsymbol{\eta}$ is beta distributed, via the geometric expectation of Result 1, we have

$$\mathsf{E}[\ln(\eta_m)] = \ln\left[\mathsf{G}(\eta_m)\right] = \Psi(\mathsf{E}[\alpha^{(m)}]) - \Psi(\mathsf{E}[\alpha^{(m)}] + D_m).$$

The variational posteriors for the auxiliary variables $\mathbf{s}$ is

$$q(\mathbf{s} \mid \mathbf{Z}) \propto \prod_{m,k} \begin{bmatrix} \hat{n}_{mk} \\ s_{mk} \end{bmatrix} (\mathsf{G}[\alpha^{(m)}\beta_k])^{s_{mk}}, \tag{6.2.11}$$

where the expectation of $\mathbf{s}$ depends on $\mathbf{Z}$ through the count $\hat{n}_{mk}$ that can take many values. Hence, the exact computation of Eq. 6.2.11 is too expensive. According to Result 3, we use the improved second-order Taylor expansion to approximate the expectation of s_{mk} as

$$\begin{aligned} \mathsf{E}[s_{mk}] \approx{}& \mathsf{G}[\alpha^{(m)}\beta_k] p_+(\hat{n}_{mk})\Big(\Psi\left(\mathsf{G}[\alpha^{(m)}\beta_k] + \mathsf{E}_+[\hat{n}_{mk}]\right) \\ &- \Psi(\mathsf{G}[\alpha^{(m)}\beta_k]) + \frac{\mathsf{V}_+[\hat{n}_{mk}]\Psi'(\mathsf{G}[\alpha^{(m)}\beta_k] + \mathsf{E}_+[\hat{n}_{mk}])}{2}\Big). \end{aligned}$$

Likewise, we can derive the variational update for **t** in the same manner. Following the exponential family computation (Wainwright and Jordan 2008), the variational updates for the remaining parameters are obtained via the conjugacy of our model specification.

Updates for the sufficient statistics: Using some algebraic manipulations, we can update the sufficient statistics as

$$\begin{aligned}
\mathsf{E}\left[\hat{n}_{mk}\right] &= \sum_d \rho_{kd}^{(m)}, \\
\mathsf{E}\left[\tilde{n}_{mk}\right] &= \sum_d (1-\rho_{kd}^{(m)}), \\
p_{+}\hat{n}_{mk}) &= 1-\exp\big(\sum_d \log[1-\rho_{kd}^{(m)}]\big), \\
p_{+}(\tilde{n}_{mk}) &= 1-\exp\big(\sum_d \log[\rho_{kd}^{(m)}]\big), \\
\mathsf{E}_{+}[\hat{n}_{mk}] &= \frac{\mathsf{E}[\hat{n}_{mk}]}{p_{+}(\hat{n}_{mk})}, \\
\mathsf{E}_{+}[\tilde{n}_{mk}] &= \frac{\mathsf{E}[\tilde{n}_{mk}]}{p_{+}(\tilde{n}_{mk})}, \\
\mathsf{V}\left[\hat{n}_{mk}\right] = \mathsf{V}\left[\tilde{n}_{mk}\right] &= \sum_d (1-\rho_{kd}^{(m)})\rho_{kd}^{(m)}, \\
\mathsf{V}_{+}[\hat{n}_{mk}] &= \frac{\mathsf{V}[\hat{n}_{mk}]}{p_{+}(\hat{n}_{mk})}, \\
\mathsf{V}_{+}[\tilde{n}_{mk}] &= \frac{\mathsf{V}[\tilde{n}_{mk}]}{p_{+}(\tilde{n}_{mk})}.
\end{aligned} \tag{6.2.12}$$

Updates for $\sigma_{w_{kd}}^{(m)}$ and $\mu_{w_{kd}}^{(m)}$: Via the normal-normal conjugacy, we update the variational parameters $\sigma_{w_{kd}}^{(m)}$ and $\mu_{w_{kd}}^{(m)}$ as

$$\sigma_{w_{kd}}^{(m)} = \left(\mathsf{E}\left[\lambda_{kd}^{(m)}\right] + \mathsf{E}\left[z_{kd}^{(m)}\right] \sum\nolimits_n \mathsf{E}\left[\tau_n^{(m)}\right] \mathsf{E}\left[f_{nk}^2\right]\right)^{-1}, \tag{6.2.13}$$

$$\mu_{w_{kd}}^{(m)} = \sigma_{w_{kd}}^{(m)} \left(\mathsf{E}\left[z_{kd}^{(m)}\right] \sum\nolimits_n \mathsf{E}\left[\tau_n^{(m)}\right] \mathsf{E}\left[f_{nk}\right] \tilde{x}_{nd}^{(m)\,-k}\right). \tag{6.2.14}$$

Updates for the auxiliary variables s, t: Exploiting Result 3, we update the auxiliary variables **s, t** as

$$\begin{aligned}
\mathsf{E}[s_{mk}] \approx\ & \mathsf{G}[\alpha^{(m)}\beta_k]p_{+}(\hat{n}_{mk})\big(\Psi\big(\mathsf{G}[\alpha^{(m)}\beta_k] + \mathsf{E}_{+}[\hat{n}_{mk}]\big) \\
& - \Psi(\mathsf{G}[\alpha^{(m)}\beta_k]) + \frac{\mathsf{V}_{+}[\hat{n}_{mk}]\Psi'(\mathsf{G}[\alpha^{(m)}\beta_k] + \mathsf{E}_{+}[\hat{n}_{mk}])}{2}\big), \\
\mathsf{E}[t_{mk}] \approx\ & \mathsf{G}[\alpha^{(m)}\bar{\beta}_k]p_{+}(\tilde{n}_{mk})\big(\Psi\big(\mathsf{G}[\alpha^{(m)}\bar{\beta}] + \mathsf{E}_{+}[\tilde{n}_{mk}]\big) \\
& - \Psi(\mathsf{G}[\alpha^{(m)}\beta_k]) + \frac{\mathsf{V}_{+}[\tilde{n}_{mk}]\Psi'(\mathsf{G}[\alpha^{(m)}\bar{\beta}] + \mathsf{E}_{+}[\tilde{n}_{mk}])}{2}\big).
\end{aligned} \tag{6.2.15}$$

Updates for $\sigma_{f_{nk}}$ and $\mu_{f_{nk}}$: Using the normal-normal conjugacy, the variational parameters $\sigma_{f_{nk}}$ and $\mu_{f_{nk}}$ can be updated as

$$\sigma_{f_{nk}} = \left(\sum\nolimits_{m,d} \mathsf{E}\left[\tau_n^{(m)}\right] \mathsf{E}\left[z_{kd}^{(m)}\right] \mathsf{E}\left[\left(w_{kd}^{(m)}\right)^2\right] + 1\right)^{-1}, \tag{6.2.16}$$

$$\mu_{f_{nk}} = \sigma_{f_{nk}} \left(\sum\nolimits_{m,d} \mathsf{E}\left[\tau_n^{(m)}\right] \mathsf{E}\left[z_{kd}^{(m)}\right] \mathsf{E}\left[w_{kd}^{(m)}\right] \tilde{x}_{nd}^{(m)\,-k}\right). \tag{6.2.17}$$

Updates for a_k and b_k:

$$\begin{aligned} a_k &= \kappa_0/K + \mathsf{E}\left[s_{\cdot k}\right], \\ b_k &= \kappa_0(1 - 1/K) + \mathsf{E}\left[t_{\cdot k}\right]. \end{aligned} \tag{6.2.18}$$

Updates for $e_{kd}^{(m)}$ and $f_{kd}^{(m)}$: Via the gamma-normal conjugacy, we have

$$\begin{aligned} e_{kd}^{(m)} &= e_0 + 1/2, \\ f_{kd}^{(m)} &= f_0 + \left(\mathsf{E}\left[\left(w_{kd}^{(m)}\right)^2\right]\right)/2. \end{aligned} \tag{6.2.19}$$

Updates for $g_n^{(m)}$ and $h_n^{(m)}$: The variational parameters $g_n^{(m)}$ and $h_n^{(m)}$ can be updated using the gamma-normal conjugacy as

$$\begin{aligned} g_n^{(m)} &= g_0 + (D_m)/2, \\ h_n^{(m)} &= h_0 + \left(\mathsf{E}\left[\|\mathbf{x}_n^{(m)} - \mathbf{G}^{(m)}\mathbf{f}_n\|^2\right]\right)/2. \end{aligned} \tag{6.2.20}$$

Updates for $c^{(m)}$ and $d^{(m)}$:

$$c^{(m)} = c_0 + \mathsf{E}\left[s_{m\cdot}\right] + \mathsf{E}\left[t_{m\cdot}\right], d^{(m)} = d_0 - \mathsf{E}\left[\log \eta_m\right]. \tag{6.2.21}$$

Updates for the auxiliary variables η: As η_m is beta distributed, we apply Result 1 and then have

$$\mathsf{E}[\log \eta_m] = \Psi(\mathsf{E}[\alpha^{(m)}]) - \Psi(\mathsf{E}[\alpha^{(m)}] + D_m). \tag{6.2.22}$$

Altogether, our CVI algorithm for the NGFA is summarized in Algorithm 7.

Algorithm 7 Collapsed variational inference for the NGFA

```
Input: data $\mathbf{X}$, model $\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{s}, \mathbf{t}, \boldsymbol{\eta})$, maximum iteraction $\mathcal{J}$, variational approximation $q(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{s}, \mathbf{t}, \boldsymbol{\eta}; \mathbf{\Phi})$, and hyper-parameter $\kappa_0$
Output: variational parameters $\mathbf{\Phi}$[1]
Initialize $\mathbf{\Phi}$ randomly.
for iter $= 1 : \mathcal{J}$ do
    for $k = 1$ to $K_+$[2] do
        Update $a_k$, $b_k$ (Eq. 6.2.18)
        for $m = 1$ to $M$ do
            Update the sufficient statistics in (Eq. 6.2.12)
            Calculate $\mathsf{E}[s_{mk}]$, $\mathsf{E}[t_{mk}]$ (Eq. 6.2.15)
            for $d = 1$ to $D_m$ do
                Update $\rho_{kd}^{(m)}$ (Eq. 6.2.10)
                Update $\sigma_{w_{kd}}^{(m)}$, $\mu_{w_{kd}}^{(m)}$ (Eq. 6.2.13; 6.2.14)
                Update $e_{kd}^{(m)}$ and $f_{kd}^{(m)}$ (Eq. 6.2.19)
            end for
        end for
        for $n = 1$ to $N$ do
            Update $\sigma_{f_{kn}}$ and $\mu_{f_{kn}}$ (Eq. 6.2.16; 6.2.17)
        end for
    end for
    for $m = 1$ to $M$ do
        Update $c^{(m)}$ and $d^{(m)}$ (Eq. 6.2.21)
        Calculate $\mathsf{E}[\log \eta_m]$ (Eq. 6.2.22)
        for $n = 1$ to $N$ do
            Update $g_n^{(m)}$ and $h_n^{(m)}$ (Eq. 6.2.20)
        end for
    end for
end for
```

6.3 EXPERIMENTS

In this section, we compare the nonparametric Bayesian group factor analysis using our proposed CVI algorithm with the state-of-the-art GFA models. We evaluate the proposed CVI algorithm on both synthetic data and real-world applications. In all our experiments, we set $\kappa_0 = 1, c_0 = 0.1, d_0 = 0.1, g_0 = 0.1, h_0 = 0.1, e_0 = 0.1, f_0 = 0.1$. Similar results are obtained when instead setting $\kappa_0 = 0.1, \kappa_0 = 10$ in a sensitivity analysis.

---

[2] For the sake of clarity, we use $\mathbf{\Phi}$ to denote all the variational parameters.

[2] We use K_+ to denote the number of active factors as the hierarchical beta Bernoulli prior can shrink the coefficients of the redundant factors to zeros.

6.3.1 *Simulated data*

For our evaluations on synthetic data, we adopt the simulation study in (Zhao et al. 2016): we performed two simulations (*Simulation 1* and *Simulation 2*) which include four groups of data with the dimensionality $D_m = 100$ for each group, respectively. The numbers of samples in the four groups are set to $N = \{20, 40, 60, 100\}$, respectively. In *Simulation 1*, we set the number of latent factors $K = 6$, and generated data only with sparse factor loadings. Specifically, the first three factors are specific to $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$, respectively, and the last three are shared among all groups. In *Simulation 2*, we set $K = 8$ and generated data with both sparse and dense factor loadings. The sparsity pattern is described in Table 6.1, and also shown in Fig. 6.3.

	Simulation 1						Simulation 2							
	1	2	3	4	5	6	1	2	3	4	5	6	7	8
$X^{(1)}$	s	-	-	s	-	-	s	-	-	-	d	-	-	-
$X^{(2)}$	-	s	-	s	s	s	-	s	-	s	-	d	-	-
$X^{(3)}$	-	-	s	-	s	s	-	-	s	s	-	-	d	-
$X^{(4)}$	-	-	-	-	-	s	-	-	s	-	-	-	-	d

Table 6.1: Sparsity pattern of the factor loading matrices in Simulation 1 and 2. “s” represents a sparse column vector; “d” represents a dense column vector; “-” represents no contribution to that group from the factor.

The sparsity of the sparse factor loadings is handled by setting 90% of the entries in each loading column to zero at random, and the nonzero entries in both the sparse and dense factor loadings are generated from a Gaussian distribution $\mathcal{N}(0, 4)$. The latent factors are generated from a standard Gaussian distribution (i.e., zero mean and unit variance). We generated the residual noise i.i.d. from a Gaussian distribution $\mathcal{N}(0, 1)$.

We compare the following methods:

1. **GFA:** The Bayesian group factor analysis model (Virtanen et al. 2012) with column-wise ARD priors to induce column-wise sparsity on the factor loading matrix. For the GFA model, we used the `GFA` package with the default parameters setting as set in the code released online. [3] The initial number of factors is set to the true values. The optimization method is L-BFGS with the maximum iterations set to 10^5.

2. **sGFA:** The extension of the GFA with element-wise ARD priors inducing element-wise sparsity (Bunte et al. 2016). For the sGFA model, the initial number of factors is set to half of the minimum of the sample size and the total number of variables, i.e., $K = \min(N, \sum_m D_m)$. The total number of MCMC iterations is set to 10^5 with sampling steps set to 10^3 and thinning steps set to 5.

3. **ssGFA:** The extension of the GFA with the spike-and-slab prior (Bunte et al. 2016), for which we again use the `GFA` package with the spike-and-slab prior. We set the noise parameters by the `informativeNoisePrior` function to prevent overfitting. The initial number of factors is set to half of the minimum of the sample size and the total number of variables. The total number of MCMC iterations is set to 10^5 with sampling steps set to 10^3 and thinning steps set to 5.

---

[3] https://cran.r-project.org/web/packages/GFA/index.html.

4. **BASS:** The Bayesian group factor analysis with structured sparsity priors (BASS) (Zhao et al. 2016), for which we use the code released in (Zhao et al. 2016). [4] The BASS is initialized using 50 iterations of MCMC and followed by expectation maximization until convergence, reached when both the number of nonzero loadings do not change for t iterations and the log-likelihood change is less than 1×10^{-5} within t iterations. The initial number of factors is set to 10 in *Simulation* 1 and 15 in *Simulation* 2 as described in (Zhao et al. 2016).

We performed 20 runs for each method, in particular to evaluate the sensitivity of our inference algorithm to initializations since CVI algorithms are only guaranteed to converge to a local optimum. For all the experiments, we simply set the initial number of factors for our method to be the minimum of the sample size and the dimensionality of each group, and ran the model with CVI algorithm until convergence.

To evaluate the performance of the methods on the recovery of sparse and dense factor loadings, we used the sparse and dense stability index defined in (Zhao et al. 2016) to quantify the distance between the true and the inferred factor loading matrices. Given the absolute correlation matrix $\mathbf{C} \in \mathbb{R}^{K_1 \times K_2}$ of the columns of two sparse matrices, the *sparse stability index* (SSI) is calculated as

$$\begin{aligned} \mathrm{SSI} = & \frac{1}{2K_1} \sum_{r=1}^{K_1} \Big(\max(\mathbf{C}_{r:}) - \frac{\sum_l \mathbb{1}(\mathbf{C}_{rl} > \hat{\mathbf{C}}_{r:}) \mathbf{C}_{rl}}{K_2 - 1} \Big) \\ & + \frac{1}{2K_2} \sum_{l=1}^{K_2} \Big(\max(\mathbf{C}_{:l}) - \frac{\sum_r \mathbb{1}(\mathbf{C}_{rl} > \hat{\mathbf{C}}_{:l}) \mathbf{C}_{rl}}{K_1 - 1} \Big), \end{aligned}$$

where $\mathbf{C}_{r:}$ and $\mathbf{C}_{:l}$ denote the r-th row and l-th column of the matrix $\mathbf{C}$, respectively; $\hat{\mathbf{C}}_{r:}$ and $\hat{\mathbf{C}}_{:l}$ denote the mean of the r-th row and l-th column of the matrix $\mathbf{C}$, respectively. The SSI is invariant to column-scaling and -permutation; larger values indicate better recovery.

The *dense stability index* (DSI) measures the distance between dense matrix columns. Given two dense matrices $\mathbf{M}_1 \in \mathbb{R}^{K_1 \times D}$ and $\mathbf{M}_2 \in \mathbb{R}^{K_2 \times D}$, the DSI is defined as

$$\mathrm{DSI} \; = \frac{1}{D^2} \mathrm{tr}(\mathbf{M}_1 \mathbf{M}_1^T - \mathbf{M}_2 \mathbf{M}_2^T).$$

The DSI is invariant to orthogonal matrix transformation, column-scaling and -permutation; the lower values indicate better recovery.

Following the strategy in (Zhao et al. 2016), in *Simulation 1* where all factor loadings are sparse, we calculated the SSI between the true and recovered factor loading matrices. In *Simulation 2*, we first thresholded the recovered factor loading matrix entries with a sparsity threshold set to 0.15. Then, we categorized the columns of each recovered factor loading matrix into sparse columns and dense columns by selecting the first 4 columns with most nonzero entries as dense columns, and the remaining columns as sparse columns. We calculated SSI between the true and the recovered sparse factor loading columns, and DSI between the true and the recovered dense columns. We calculated the two stability indices for each group separately and averaged the result for all groups.

---

[4] https://github.com/judyboon/BASS.

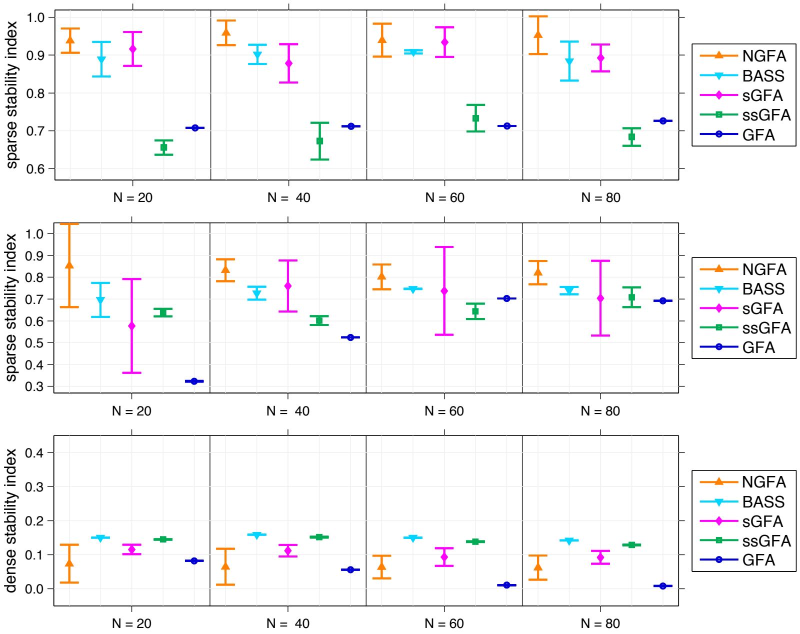


Figure 6.2: The comparison of stability indices on the inferred matrix of factor loadings for our synthetic data. For SSI, higher is better; for DSI, lower is better. The means and the standard derivations of the stability indices are denoted by the marker and the bar respectively. The SSI comparisons of all the methods in Simulation 1 are shown in upper rows; The SSI and DSI comparisons in Simulation 2 are shown in middle and bottom rows, respectively.

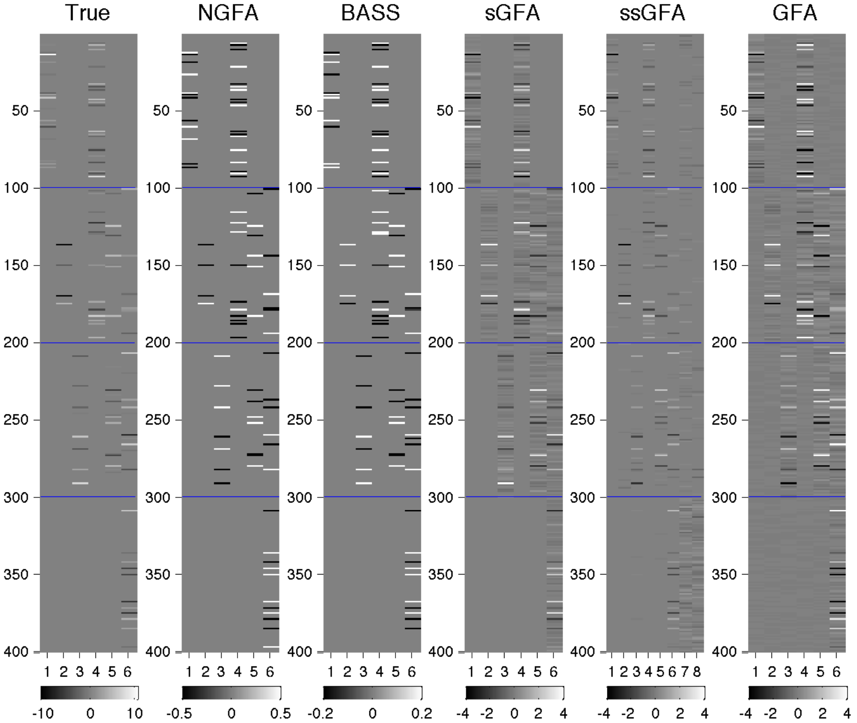


Figure 6.3: The true and the inferred factor loadings by all the methods in *Simulation 1*. The columns of the inferred factor loading matrices were reordered for easy comparison. The horizontal lines separate the four groups.

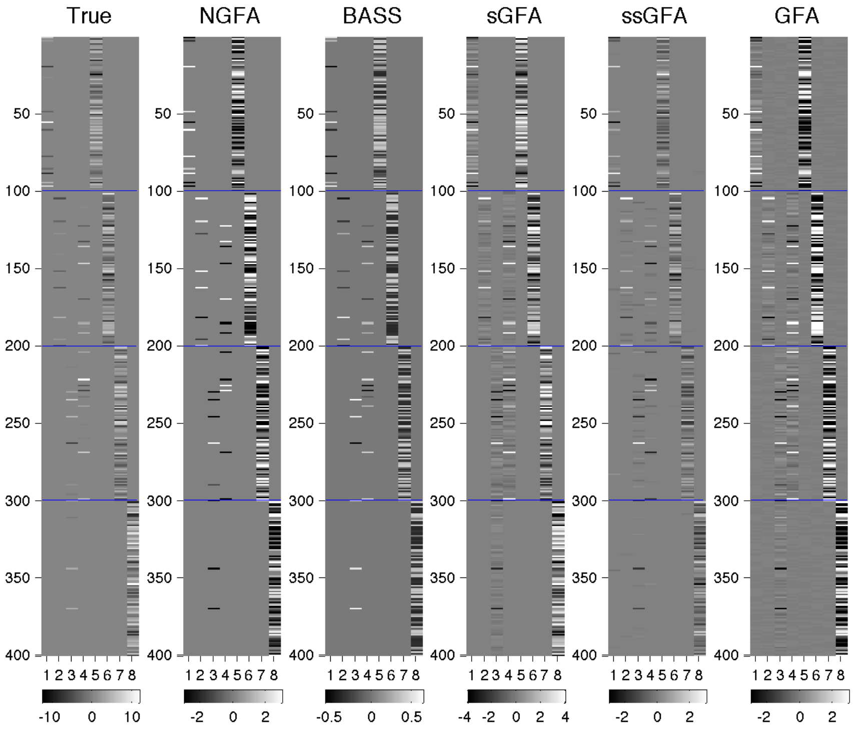


Figure 6.4: The true and the inferred factor loadings by all the methods in *Simulation 2*. The columns of the inferred factor loading matrices were reordered for easy comparison. The horizontal lines separate the four groups.

The true and the inferred factor loading matrices by all methods in *Simulation 1* and *Simulation 2* are shown in Fig. 6.3. The ARD prior cannot induce sufficient sparsity by pushing irrelevant factor loadings to small values. As a consequence, the GFA has difficulty in recovering sparse factor loadings because of the columns-wise ARD priors (Fig. 6.3). Similarly, the sGFA cannot induce sufficient element-wise sparsity within the loading columns by the independent ARD priors (Fig. 6.3). The ssGFA overfitted to data by not sufficiently shutting off the redundant factors (Fig. 6.3). Both the BASS and NGFA achieve element-wise sparsity effectively (Fig. 6.3). We quantified the performance of the methods with stability indices, i.e., the means and the standard derivations of the stability indices for each method over 20 runs are shown in Fig. 6.2. The NGFA using our CVI algorithm achieves the best SSI and DSI scores almost for all sample sizes.

6.3.2 *Cancer gene prioritization*

Integrative analysis of multiple genomic datasets for understanding the genetic basis of common diseases has been challenging. For instance, DNA alterations that are frequent in cancers, measured by copy number variation (CNV) data, are known to induce gene expression modifications. Hence, cancer-related genes can be discovered by searching for such interactions. Recently, Bayesian GFA methods were applied to the task of cancer gene prioritization with encouraging results (Klami, Virtanen, and Kaski 2013). To demonstrate the effectiveness of the NGFA using our CVI algorithm, we choose the same datasets `Hyman` and `Pollack` from (Lahti et al. 2013) that are based on gene expression (GE) and CNV data as described in Table 6.2.

Dataset	# genes	# samples	# cancer genes
Hyman	7489	14	48
Pollack	4287	41	38

Table 6.2: The details of the cancer genomics datasets.

More specifically, we consider the patients as co-occurring samples and all the genes in the whole genome as features. The GE and CNV data constitute the two groups. We then rank the genes according to the quantity defined by $s_d = \sum_{k=1}^{K} |\mathsf{E}(g_{kd}^{(1)})\mathsf{E}(g_{kd}^{(2)})|$, that is the correlation between GE and CNV data captured by the shared factors. We repeated the data pre-processing procedure in (Lahti et al. 2013), and evaluated the model performance by the area under the curve of the receiver operating characteristic (AUROC) for retrieving known cancer-related genes. We ran the NGFA 20 times with the initial K set to the minimum of the sample size and feature dimension. We compared the NGFA using CVI algorithm to the Bayesian inter-battery factor analysis (BIBFA) model. We ran the BIBFA for 20 times according to the setting described in (Klami, Virtanen, and Kaski 2013). The mean AUROC scores and the standard deviations are shown in Fig. 6.5. The AUROC scores for all the other methods are cited from (Lahti et al. 2013) where the standard deviations cannot be presented because those alternatives are deterministic methods. The NGFA using our CVI algorithm outperforms all the alternative methods.

6.3.3 *Decoding fMRI brain activity*

Bayesian canonical correlation analysis (BCCA) was investigated to analyze fMRI responses to visual stimuli in (Fujiwara, Miyawaki, and Kamitani 2009). We evaluated the NGFA using our

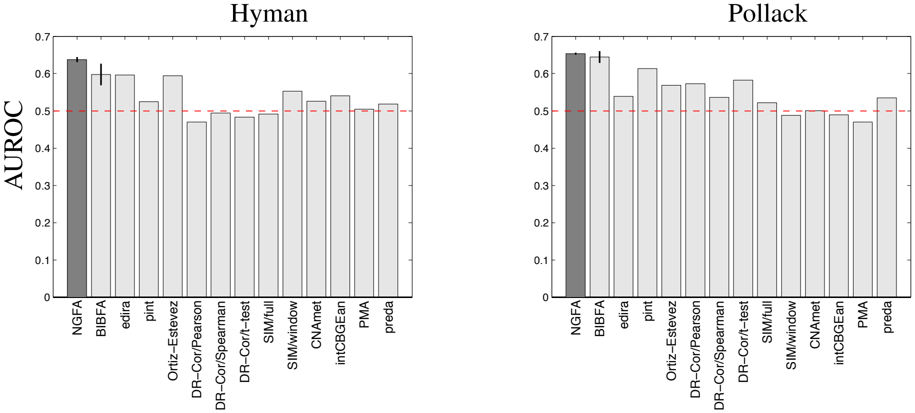


Figure 6.5: Evaluation of cancer gene prioritization performance of the methods on two data sets: Hyman (left) and Pollack (right). The result is quantified by the area under the ROC curve (AUROC). The dashed line indicates the AUROC score for a random list (AUROC = 0.5). The comparison shows that the NGFA achieves best performance.

CVI algorithm to the fMRI recordings of two subjects viewing visual images consisting of contrast-defined 10×10 patches (Miyawaki et al. 2008). The data is composed of two independent sessions: one for "random image session" with spatially random patterns sequentially presented; the other for "figure image session" with alphabet letters and geometric shapes sequentially presented. For the NGFA, we first treated the random image session and corresponding fMRI recordings as two groups, to extract the image bases and weight vector automatically from the input, with the initial K set to $\min(D_1, D_2) = 100$. Our task is to reconstruct the visual image from the new fMRI recordings in the figure image session. The reconstruction performance is evaluated by the mean squared error between the presented and reconstructed images. We ran both the BCCA and the NGFA for 20 times. The mean squared prediction error over 20 runs for NGFA is 0.224 with the standard deviation less than 1e-3, which is better than the result 0.251(0.002) of the BCCA. The reconstructed geometric shapes and alphabet letters by the BCCA and the proposed NGFA are shown in Fig. 6.6.

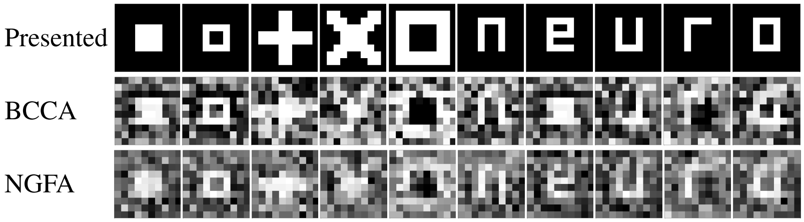


Figure 6.6: Presented images (first row) and the reconstructed visual images obtained from the BCCA (second row) and the NGFA (third row).

6.4 CONCLUSIONS

In this chapter, the group factor analysis problem is tackled via a Bayesian nonparametric method that allows the total number of factors to be automatically inferred, and the underlying structured sparsity to be effectively captured. In particular, we have presented an efficient collapsed variational inference algorithm for the nonparametric Bayesian group factor analysis model. By integrating out the group-specific beta process parameters, the proposed collapsed variational inference algorithm achieves a better approximation because all latent variables are dependent through the field while the weak dependences are very small in the collapsed space. Using the Gaussian approximation technique, all the variational parameters can be efficiently updated through closed form expressions. Experimental results on both synthetic data and real-world applications demonstrate the superior performance of the proposed CVI algorithm for the nonparametric Bayesian group factor analysis model when compared to state-of-the-art GFA methods.

7

JOINT MODELS FOR NETWORK EDGES AND NODE FEATURES

Analysis of relational data is becoming an increasingly important problem in many domains such as computational system biology (Hill et al. 2012; Oates et al. 2014) and social network analysis (Hamilton et al. 2017). On the one hand, the scope and availability of relational data arising from these domains increases. For instance, high-throughput methods for screening protein-protein interactions (PPIs) enable us to characterize large-scale protein interaction networks (Zitnik et al. 2019). As the rapid growth of internet technology, networks arising from online platforms like Facebook and Twitter are exponentially growing. On the other aspect, these collected networks are incomplete with edges and nodes missing. In recent years, there is a growing interest in leveraging the available node features to reconstruct the missing edges with successful applications in recommending new friends in social networks or new items to users in recommender systems.

Many previous network models (Chang and D. Blei 2009; M. Kim and Leskovec 2012; D. I. Kim et al. 2012; Rai 2017) have been studied to either jointly model the network structure and its associated node features, or leverage the node features into the prediction of missing edges using a regression approach. For example, the relational topic models (RTMs) (Chang and D. Blei 2009) are developed to generalize the latent Dirichlet allocation (D. M. Blei et al. 2003) to model both the underlying topics behind documents and the links among these documents simultaneously. Extending relational topic models for discrete-count data, the nonparametric Metadata dependent relational model (NMDR) (D. I. Kim et al. 2012) can model both discrete and continuous node-specific metadata, and determines the number of latent communities in a Bayesian nonparametric way. Another line of research is to treat the node features as covariates, and to predict the missing edges using a regression-based approach. For instance, the inductive latent factor model (ILFM) of (Rai 2017) generalizes the gamma process edge partition model (M. Zhou et al. 2015) to impute missing edges using available node features. Instead of modeling the node features, the ILFM incorporates the node features via the scale parameter of the gamma distributed node-community memberships. Despite performing well in various network completion scenarios, the ILFM scales poorly to the network data with large dimensional node features. In this chapter, we generalize the single Poisson gamma memberships framework to jointly model the given network data and its associated nonnegative real-valued node features.

7.1 THE POISSON GAMMA MEMBERSHIPS MODEL FOR NETWORK EDGES AND NODE FEATURES

The proposed joint Poisson gamma memberships model generates the network structure and its associated node features. Let $A \in \{0,1\}^{V \times V}$ be the adjacency matrix of V nodes, and each node $u \in \mathcal{V}$ is associated with a nonnegative real-valued feature vector $\mathbf{x}_u \in \mathbb{R}^D$. In the proposed model, each node u is endowed with two gamma distributed latent node memberships vectors $\boldsymbol{\phi}_u$

and $\boldsymbol{\psi}_u$, which interpret the underlying structure in the observed network and the node features, respectively. Formally, we draw the memberships $\boldsymbol{\phi}_u$ and $\boldsymbol{\psi}_u$ as

$$\begin{aligned}\phi_{uk} &\sim \text{Gamma}(\xi_u, \frac{1}{c}), &(7.1.1)\\ \xi_d &\sim \text{Gamma}(e_0, \frac{1}{f_0}),\\ \psi_{ul} &\sim \text{Gamma}(\alpha_u, \frac{1}{d}),\\ \alpha_u &\sim \text{Gamma}(e_0, \frac{1}{f_0}),\end{aligned}$$

where the parameter ξ_u measures the overall popularity of node u irrespective of its memberships to the multiple communities. Similarly, the parameter α_u measures the degree of node u in the observed features. The proposed model assumes that the observed network is composed of K overlapping latent communities. We generate a nonnegative community weight r_k for each latent community as

$$r_k \sim \text{Gamma}(\frac{\gamma_0}{K}, \frac{1}{f}). \tag{7.1.2}$$

For fixed γ_0, the redundant latent communities can be automatically shrunk as the number of communities K increases to infinity, and then the weights of the redundant communities tend to be zeros. On the other side, a Poisson latent factor model is employed to factorize the nonnegative real-valued node features via the Poisson randomized gamma distribution as

$$x_{ud} \sim \text{PRG}(\sum_{l=1}^{L} \rho_l \psi_{ul} \beta_{dl}, \frac{1}{e_d}), \tag{7.1.3}$$

where we utilize $L < D$ latent factors to represent the correlation structure in node features; the variance-to-mean ratio of the PRG distribution is controlled by e_d. Each latent factor $\boldsymbol{\beta}_d$ is drawn from a gamma distribution as

$$\beta_{dl} \sim \text{Gamma}(\eta_d, \frac{1}{h}). \tag{7.1.4}$$

As we place a gamma prior over ρ_l as

$$\rho_l \sim \text{Gamma}(\frac{\tau_0}{L}, \frac{1}{g}), \tag{7.1.5}$$

where the number of latent factors can be automatically inferred as we did in determining the number of latent communities K.

Finally, we generate the observed edge between a pair of nodes u and v as

$$\begin{aligned}A_{uv} &= \mathbb{1}(\tilde{A}_{uv} \geq 1), &(7.1.6)\\ \tilde{A}_{uv} &\sim \text{Poisson}(\sum_{k=1}^{K} r_k \phi_{uk} \phi_{vk} + \sum_{l=1}^{L} \rho_l \psi_{ul} \psi_{vl}).\end{aligned}$$

The intuition behind Eq. 7.1.6 is that the term $\sum_{k=1}^{K} r_k \phi_{uk} \phi_{vk} + \sum_{l=1}^{L} \rho_l \psi_{ul} \psi_{vl}$ jointly captures the edge probability between nodes u and v, and the correlation structure in their node features when

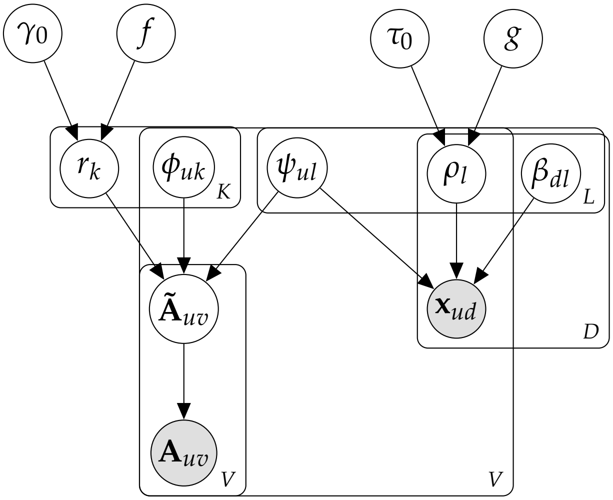


Figure 7.1: The plate notation of the joint Poisson gamma memberships model for network edges and node features (hyperparameters not shown for brevity).

an edge is present between nodes u and v. If the observation of an edge A_{uv} is missing between nodes u and v while these two nodes are highly correlated in feature data, the term $\sum_{l=1}^{L} \rho_l \psi_{ul} \psi_{vl}$ will capture the probability of the missing edge.

The full generative model is as follows

$$
\begin{aligned}
\phi_{uk} &\sim \text{Gamma}(\xi_u, \frac{1}{c}), \\
\xi_d &\sim \text{Gamma}(e_0, \frac{1}{f_0}), \\
\psi_{ul} &\sim \text{Gamma}(\alpha_u, \frac{1}{d}), \\
\alpha_u &\sim \text{Gamma}(e_0, \frac{1}{f_0}), \\
\beta_{dl} &\sim \text{Gamma}(\eta_d, \frac{1}{h}), \\
r_k &\sim \text{Gamma}(\frac{\gamma_0}{K}, \frac{1}{f}), \\
\rho_l &\sim \text{Gamma}(\frac{\tau_0}{L}, \frac{1}{g}), \\
x_{ud} &\sim \text{PRG}(\sum_{l=1}^{L} \rho_l \psi_{ul} \beta_{dl}, \frac{1}{e_d}), \\
\tilde{A}_{uv} &\sim \text{Poisson}(\sum_{k=1}^{K} r_k \phi_{uk} \phi_{vk} + \sum_{l=1}^{L} \rho_l \psi_{ul} \psi_{vl}), \\
A_{uv} &= \mathbb{1}(\tilde{A}_{uv} \geq 1).
\end{aligned}
\tag{7.1.7}
$$

Fig. 7.1 presents the plate notation of the joint Poisson gamma memberships model.

7.2 INFERENCE

The proposed model admits a full local conjugate inference scheme using the data augmentation and marginalization technique. The inference procedure requires the sampling of the model parameters including $\{\tilde{A}_{uv}, \boldsymbol{\phi}_u, \boldsymbol{\psi}_u, \boldsymbol{\beta}_d, \rho_l, r_k, \xi_u, \alpha_u, \gamma_0, \tau_0, e_d\}$.

Sampling the latent count $\tilde{A}_{uv}$: We sample the latent count $\tilde{A}_{uv}$ as

$$(\tilde{A}_{uv} \mid -) \sim A_{uv}\text{Poisson}_{+}\left(\sum_{k=1}^{K} r_k \phi_{uk}\phi_{vk} + \sum_{l=1}^{L} \rho_l \psi_{ul}\psi_{vl}\right). \tag{7.2.1}$$

Sampling the latent count $\tilde{x}_{ud}$: We sample the latent count $\tilde{x}_{ud}$ as

$$(\tilde{x}_{ud} \mid -) \sim \text{Bessel}_{-1}\left(2\sqrt{e_d x_{ud} \sum_{l=1}^{L} \rho_l \psi_{ul}\beta_{ld}}\right). \tag{7.2.2}$$

Then, we sample the latent sub counts $\tilde{x}_{udl}$ as

$$(\{\tilde{x}_{udl}\}_l \mid -) \sim \text{Multinomial}\left(\tilde{x}_{ud}; \frac{\rho_l \psi_{ul}\beta_{ld}}{(\sum_{l=1}^{L} \rho_l \psi_{ul}\beta_{ld})}\right). \tag{7.2.3}$$

Sampling the latent subcount $\tilde{A}_{uvk}$ and $\hat{A}_{uvl}$: We sample the latent subcount $\tilde{A}_{uvk}$ and $\hat{A}_{uvl}$ as

$$(\{\tilde{A}_{uvk}\}_k, \{\hat{A}_{uvl}\}_l \mid -) \sim \text{Multinomial}\left(\tilde{A}_{uv}; \frac{\{\{r_k\phi_{uk}\phi_{vk}\}_k, \{\rho_l\psi_{ul}\psi_{vl}\}_l\}}{(\sum_{k=1}^{K} r_k\phi_{uk}\phi_{vk} + \sum_{l=1}^{L} \rho_l\psi_{ul}\psi_{vl})}\right) \tag{7.2.4}$$

Sampling the community weights $\{r_k\}$: Using the gamma-Poisson conjugacy, we sample r_k as

$$(r_k \mid -) \sim \text{Gamma}\left[\frac{\gamma_0}{K} + \tilde{A}_{\cdot\cdot k}, \frac{1}{c + \sum_{u,v\neq u} \phi_{uk}\phi_{vk}}\right], \tag{7.2.5}$$

where $\tilde{A}_{\cdot\cdot k} \equiv \sum_{u,v\neq u} \tilde{A}_{uvk}$.

Sampling the node-community memberships Φ: We sample ϕ_{uk} via the gamma-Poisson conjugacy as

$$(\phi_{uk} \mid -) \sim \text{Gamma}\left[\xi_u + \tilde{A}_{uk\cdot}, \frac{1}{c + \sum_{v\neq u} r_k\phi_{vk}}\right]. \tag{7.2.6}$$

where $\tilde{A}_{uk\cdot} \equiv \sum_{v\neq u} \tilde{A}_{ukv}$.

Sampling the factor weights $\{\rho_l\}$: The factor weights $\{\rho_l\}_l$ can be sampled using the gamma-Poisson conjugacy as

$$(\rho_l \mid -) \sim \text{Gamma}\left[\frac{\tau_0}{L} + \hat{A}_{\cdot\cdot l} + \tilde{x}_{\cdot\cdot l}, \frac{1}{d + \sum_{u,v\neq u} \psi_{ul}\psi_{vl} + \sum_{u,d} \psi_{ul}\beta_{dl}}\right], \tag{7.2.7}$$

Sampling the factor loadings Ψ: We sample $\boldsymbol{\psi}_u$ via the gamma-Poisson conjugacy as

$$(\psi_{ul} \mid -) \sim \text{Gamma}\left[\alpha_u + \hat{A}_{ul\cdot} + \tilde{x}_{ul\cdot}, \frac{1}{c + \sum_{v\neq u} \rho_l\psi_{vl} + \sum_d \rho_l\beta_{dl}}\right]. \tag{7.2.8}$$

where $\hat{A}_{ul\cdot} \equiv \sum_{v \neq u} \hat{A}_{ulv}$.

Sampling the factors $\boldsymbol{\beta}_d$: We sample $\boldsymbol{\beta}_d$ via the gamma-Poisson conjugacy as

$$(\beta_{dl} \mid -) \sim \text{Gamma}\Big[\eta_d + \tilde{x}_{\cdot dl}, \frac{1}{h + \sum_u \rho_l \psi_{ul}}\Big]. \tag{7.2.9}$$

where $\hat{A}_{ul\cdot} \equiv \sum_{v \neq u} \hat{A}_{ulv}$.

We exploit the gamma-Poisson conjugacy to derive the posterior distribution of parameters γ_0, τ_0, f, g.

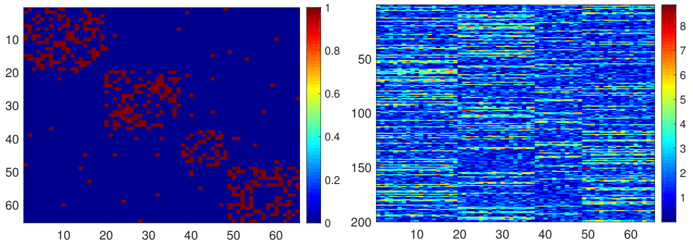
7.3 EXPERIMENTS

In this section, we demonstrate the developed joint Poisson gamma memberships model for network reconstruction on synthetic data. The synthetic data were generated as follows. We considered $V = 65$ nodes, $K = 4$ latent communities, and generated the intra-community edges $A_{uv} \sim \text{Bernoulli}(\Pi_{uv})$, where $\Pi_{uv} \sim \text{Beta}(0.7, 0.7)$, and the inter-community edges $A_{uv} \sim \text{Bernoulli}(\Pi_{uv})$, where $\Pi_{uv} \sim \text{Beta}(0.2, 0.2)$. To generate the synthetic node features $\mathrm{X} \in \mathbb{R}^{V \times D}$, three datasets were generated with the number of samples $D = 200, 500, 1000$, respectively. We assume that the nodes affiliated with the same communities are correlated in their corresponding node features. Hence, we generated the node features using the same covariance matrix for the nodes affiliated with the same communities. One of the simulated networks and its associated node features of $D = 200$ samples are shown in Fig. 7.2. To evaluate the performance of network reconstruction, we randomly held out a fraction $\rho = 0.2, 0.4, 0.6, 0.8$ of the network entries (include both zeros and non-zeros) as test data, and used the remaining network entries as the training data. We compared the joint Poisson gamma memberships model (joint-PGMM) to the single Poisson gamma memberships (PGMM) model that only accounts for the network strucutre.

For the two models, we ran 2000 burn-in MCMC iterations, and collected last 1000 samples from the model posterior distribution. The posterior mean of the edge probability is estimated for each held-out edge in the test data by averaging over the collected Gibbs samples. These edge probabilities were used to evaluate the predictive performance of each model by calculating the area under the curve of the receiver operating characteristic (AUROC) and of the precision-recall (PR). Fig. 7.3 presents the results. Overall, we found that the joint-Poisson gamma memberships model performs better than single Poisson gamma memberships model in network reconstruction as the joint-PGMM leverages the available node features, and thus better captures the missing edges. As the number of the available samples is increasing, the joint-PGMM can effectively leverage correlations in node features into estimating missing edges, thus achieves better performance.

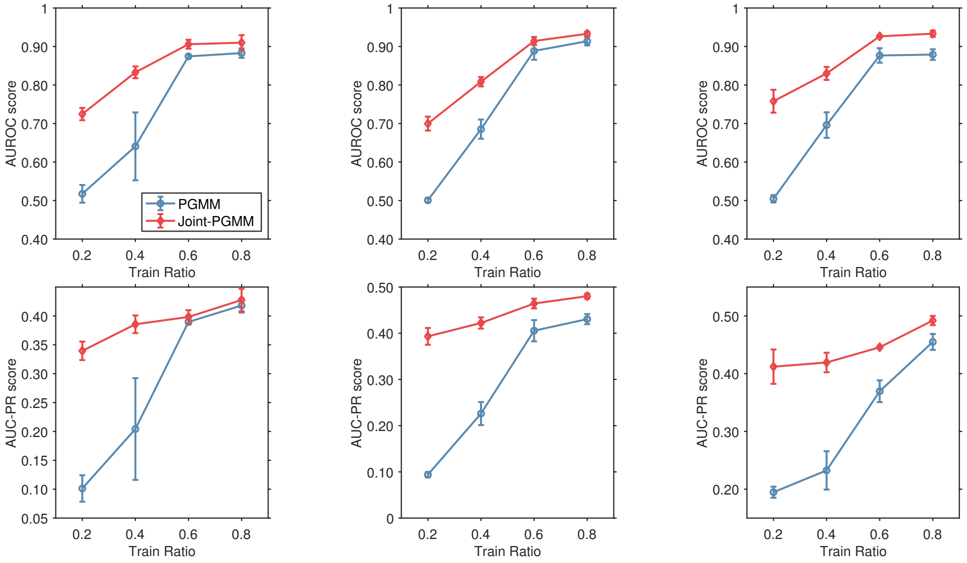
7.4 CONCLUSIONS

In this chapter, we developed a joint Poisson gamma memberships (joint-PGMM) framework for modelling the observed network and its associated node features. The joint-PGMM models the adjacency matrix of the observed network with the basic bilinear Poisson factorization component, and factorizes the positive real-valued node features using another Poisson latent factor model. This framework can leverage the available node features into estimating the missing edges using the shared node memberships vector. An efficient Gibbs sampling scheme is developed to perform posterior simulation. The experimental results on a simulated dataset demonstrate that the model capacity of the joint-PGMM, compared with a single Poisson gamma memberships model.



(a) (b)

Figure 7.2: One of the simulated network (a) and its associated node features (b).



(a) (b) (c)

Figure 7.3: Network reconstruction on synthetic data. We considered the number of samples $D = 200, 500, 1000$ as shown in column (a-c), respectively. We randomly held out a fraction $\rho = 0.2, 0.4, 0.6, 0.8$ of the network entries (include both zeros and non-zeros) as test data.

8

CONCLUSIONS AND FUTURE RESEARCH

8.0.1 *Summary of our contributions*

This thesis has developed a family of probabilistic models and inference algorithms to achieve the following objectives:

- To develop a Bayesian nonparametric model that captures the underlying overlapping community structure and characterizes the evolving node-community relationships in discrete-time dynamic networks.
- To develop a probabilistic model that not only characterizes the reciprocating interactions among nodes in continuous-time event-based networks, but also accounts for the latent structure underlying the observed interactions.
- To develop scalable inference algorithms for the proposed dynamic network framework.
- To develop a unified framework to jointly model the common structure and group-specific signals in multiple related groups of data.

In this thesis, the dynamic Poisson gamma memberships model is proposed in Chapter 3 to achieve the first objective. The proposed model represents each node using a gamma-distributed memberships vector that effectively captures the underlying overlapping community structure. For discrete-time networks, the dynamic node-community relationships are captured by evolving node memberships via gamma Markov processes. Moreover, we utilize a community-community interaction matrix built upon the hierarchical gamma process to characterize both intra-community and inter-community interactions between nodes. In particular, the number of latent communities can be automatically determined by the shrinkage mechanism of the hierarchical gamma process in a Bayesian nonparametric way. Using the Bernoulli Poisson link function that maps binary network edges into latent space, the proposed models are suitable to fit sparse dynamic networks. Because the inference only needs to be performed on non-zero edges, the proposed model is significantly less computational demanding, compared with the probabilistic models using the probit/logistic link function. Furthermore, we also exploit the time-dependent hierarchical gamma process to capture the birth and death dynamics of latent communities, which enables us to analyze and understand the formation and decaying processes of latent communities.

In Chapter 4, the Hawkes edge partition model (Hawkes-EPM) is developed to capture the overlapping community structure underlying the timestamped interaction events among nodes. To capture the reciprocating behaviour in continuous-time networks, the inferred community structure is incorporated into the base intensity of the mutually-exciting Hawkes process for each pair of two nodes to characterize the exogenous interaction events between these two nodes. The proposed model augments each interaction event between a pair of two nodes with a pair of latent variables, to indicate which of their latent communities (features) leads to the occurring of that interaction. Moreover, this model allows the excitation effect of each interaction on its opposite direction is determined by its latent variables. Both Gibbs sampling with closed-form update equations and

Expectation-Maximization algorithms are derived to perform inference for the proposed Hawkes-EPM.

Although the developed Gibbs sampling scheme for performing inference in the proposed models are simple yet efficient, we also exploit the recently advanced stochastic gradient Riemannian Langevin dynamics algorithms to further scale up the derived inference procedures in Chapter 5.

Moreover, a Bayesian nonparametric group factor analysis model is investigated in Chapter 6 to estimate the common structure and group-specific signals from multiple-related groups of data. The proposed model factorizes the multi-related matrices using a set of common factors shared among the observed multi-groups, and then reconstructs each group with a group-specific matrix of factor loadings. To improve the model flexibility, the hierarchical beta-Bernoulli process is investigated to induce sparsity over the factor loading matrices. To adapt the proposed method for modelling large-scale data, a collapsed variational Bayesian algorithm is developed to perform inference. Compared with state-of-the-art group factor analysis methods, the proposed model demonstrates improved predictive performance and highly interpretable parameters.

Finally, a probabilistic framework for joint modelling of the observed network structure and its associated node features is investigated in Chapter 7. The simulated example shows the mechanism of the proposed framework that can be utilized to reconstruct missing network edges using available node features.

8.0.2 *Future Work*

First, the current statistical methods for network inference can discover the underlying properties of entities, and predict missing edges using observed network entries and available node features, which enable us to precisely profile individuals and to recommend new links on social networks. As the scope and availability of social interaction data are increasing, concerns about the privacy of these data have become an increasingly important issue. Hence, there is a growing need to develop privacy-preserving network analysis methods. Recently, a privacy-preserving Bayesian inference scheme has been developed for Poisson factorization methods (Schein et al. 2018). An interesting direction of research is to extend the privacy-preserving inference methods for modelling dynamic networks.

Another interesting direction of extension is to infer the time-evolving hierarchical community structure revealed by dynamic network data. Most current network models factorize the observed dynamic network into *flat* latent community structure, which may not be able to sufficiently interpret the real-world complicated interaction data. For instance, the community of computer scientists can be split into those working on various branches of computer science like algorithm design and artificial intelligence, and each branch repeatedly split until reaching the particular research topic of an individual scientist. For dynamic context, such tree-structured latent community hierarchies are growing and evolving over time. Hence, it is necessary to develop models that can represent hierarchically-organized entities with multi-layers of latent node-community memberships. Other extensions of current work include modelling and reasoning of dynamic knowledge graphs (Nickel et al. 2016) via the bilinear Poisson factorization framework. Knowledge graphs represent multi-relationships among entities. As the availability of large-scale temporal event data where each edge is associated with a timestamp, it is necessary to build models that embeds entities and timing edges among them with interpretable latent parameters.

ACRONYMS

ARD	Automatic Relevance Determination
AUROC	Area Under the curve of the Receiver Operating Characteristic
BCS	Bioinspired Communication Systems
BNP	Bayesian Nonparametric Prior
BPL	Bernoulli-Poisson Link
BNHP	Bayesian Nonparametric Hawkes Process
BASS	Bayesian group factor Analysis with Structured Sparsity
BCCA	Bayesian Canonical Correlation Analysis
BCDF	Bayesian Conditional Density Filtering
CRT	Chinese Restaurant Table
CRM	Completely Random Measure
CVI	Collapsed Variational Inference
CTMC	Continuous Time Markov Chain
CCRM	Compound Completely Random Measure
CGS	Collapsed Gibbs Sampling
CNV	Copy Number Variation
tCRMs	thinned Completely Random Measures
DRIFT	Dynamic Relational Infinite Feature Model
DEPM	Dynamic Edge Partition Model
DTMC	Discrete Time Markov Chain
D-GPPF	Dynamic Gamma Process Poisson Factorization
DMMG	Dynamic Multi-group Membership Graph
DRGPM	Dynamic Relational Gamma Process Model
DSBM	Dynamic Stochastic Block Model
DPGM	Dynamic Poisson Gamma Memberships
DP	Dirichlet Process

DHP	Dirichlet Hawkes Process
DLS	Dual Latent Space
DSI	Dense Stability Index
dSBM	dynamic Stochastic Block Model
dLFRM	dynamic Latent Feature Relational Model
dRGaP	dynamic Relational Gamma Process
dd-IBP	distance dependent-Indian Buffet Process
EM	Expectation Maximization
EKF	Extended Kalman Filter
EPM	Edge Partition Model
ER	Erdös-Rényi
ERGM	Exponential Random Graph Model
EU	European Union
ELBO	Evidence Lower BOund
FA	Factor Analysis
FFDC	Face-to-Face Dynamic Contact
GFA	Group Factor Analysis
GE	Gene Expression
GaP	Gamma Process
HDP	Hierarchical Dirichlet Process
HBP	Hierarchical Beta Process
HGP	Hierarchical Gamma Process
HGPEPM	Hierarchical Gamma Process Edge Partition Model
HP	Hawkes Process
IBP	Indian Buffet Process
IRM	Infinite Relational Model
LDA	Latent Dirichlet Allocation
ILFM	Inductive Latent Factor Model
LFRM	Latent Feature Relational Model

LFP	Latent Feature Propagation
MCMC	Markov Chain Monte Carlo
MMSBM	Mixed Memberships Stochastic Block Model
MID	Military Interstate Disputes
NB	Negative-Binomial
NGFA	Nonparametric Bayesian Group Factor Analysis
NATO	North Atlantic Treaty Organization
NCRP	Nested Chinese Restaurant Process
NMDR	Nonparametric Metadata Dependent Relational model
NIG	Normal Inverse Gamma
NIPS	Neural Information Processing System
PP	Poisson Process
PR	Precision Recall
PRG	Poisson Randomized Gamma
PPI	Protein-Protein Interaction
PRG	Poisson Randomized Gamma
PGMM	Poisson Gamma Memberships Model
RTM	Relational Topic Model
SBM	Stochastic Block Model
SSI	Sparse Stability Index
SG-MCMC	Stochastic Gradient Markov Chain Monte Carlo
SGRLD	Stochastic Gradient Riemannian Langevin Dynamics
sGFA	sparse Group Factor Analysis
ssGFA	spike-and-slab Group Factor Analysis
tGaP	thinned Gamma Process

BIBLIOGRAPHY

Acharya, Ayan et al. (2015a). “Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1–9 (cit. on p. 16).

Acharya, Ayan et al. (2015b). “Nonparametric Dynamic Network Modeling”. In: *SIGKDD Workshop on Mining and Learning from Time Series*, pp. 104–113 (cit. on pp. 16, 30).

Airoldi, Edoardo M. et al. (2008). “Mixed Membership Stochastic Blockmodels”. *Journal of Machine Learning Research* 9, pp. 1981–2014 (cit. on pp. 11, 16).

Asur, Sitaram et al. (2009). “An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs”. *ACM Trans. Knowl. Discov. Data* 3.4, 16:1–16:36 (cit. on p. 39).

Blackwell, David and James B. MacQueen (1973). “Ferguson Distributions Via Polya Urn Schemes”. *Ann. Statist.* 1.2, pp. 353–355 (cit. on p. 8).

Blei, David M. et al. (2003). “Latent Dirichlet Allocation”. *Journal of Machine Learning Research* 3, pp. 993–1022 (cit. on pp. 75, 95).

Blei, David M. et al. (2010). “The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies”. *Journal of the ACM* 57.2, 7:1–7:30 (cit. on p. 13).

Blundell, Charles, Jeff Beck, and Katherine A Heller (2012). “Modelling Reciprocating Relationships with Hawkes Processes”. In: *Advances in Neural Information Processing Systems 25*, pp. 2600–2608 (cit. on pp. 10, 12, 47, 48).

Bunte, Kerstin, Eemeli Leppäaho, Inka Saarinen, and Samuel Kaski (2016). “Sparse group factor analysis for biclustering of multiple data sources”. *Bioinformatics* 32.16, pp. 2457–2463 (cit. on pp. 75, 86).

Carvalho, Carlos M. et al. (2008). “High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics”. *Journal of the American Statistical Association* 103.484, pp. 1438–1456 (cit. on p. 75).

Chang, Jonathan and David Blei (2009). “Relational Topic Models for Document Networks”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 81–88 (cit. on p. 95).

Chen, Bo et al. (2011). “The Hierarchical Beta Process for Convolutional Factor Analysis and Deep Learning”. In: *International Conference on Machine Learning*, pp. 361–368 (cit. on p. 75).

Chen, Changyou et al. (2016). “Stochastic Gradient MCMC with Stale Gradients”. In: *Advances in Neural Information Processing Systems*, pp. 2937–2945 (cit. on p. 63).

Chen, Tianqi et al. (2014). “Stochastic Gradient Hamiltonian Monte Carlo”. In: *International Conference on Machine Learning*, pp. 1683–1691 (cit. on p. 63).

Cong, Yulai et al. (2017). “Fast Simulation of Hyperplane-Truncated Multivariate Normal Distributions”. *Bayesian Analysis* 12.4, pp. 1017–1037 (cit. on pp. 69, 70).

Daley, D. J. et al. (2003). *An Introduction to the Theory of Point Processes*. Second. New York: Springer-Verlag (cit. on p. 10).

Ding, Nan et al. (2014). “Bayesian Sampling Using Stochastic Gradient Thermostats”. In: *Advances in Neural Information Processing Systems*, pp. 3203–3211 (cit. on p. 63).

Du, Nan et al. (2015a). “Dirichlet-Hawkes Processes with Applications to Clustering Continuous-Time Document Streams”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 219–228 (cit. on pp. 47, 49).

Du, Nan et al. (2015b). “Time-Sensitive Recommendation From Recurrent User Activities”. In: *Advances in Neural Information Processing Systems*, pp. 3492–3500 (cit. on pp. 12, 47).

DuBois, Christopher et al. (2013). “Stochastic blockmodeling of relational event dynamics”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 238–246 (cit. on p. 47).

Dunlavy, Daniel M. et al. (2011). “Temporal Link Prediction Using Matrix and Tensor Factorizations”. *ACM Trans. Knowl. Discov. Data* 5.2, 10:1–10:27 (cit. on pp. 1, 11, 16).

Dunson, David B. et al. (2005). “Bayesian latent variable models for mixed discrete outcomes”. *Biostatistics* 6.1, pp. 11–25 (cit. on pp. 15, 65).

Durante, Daniele et al. (2014). “Bayesian Logistic Gaussian Process Models for Dynamic Networks”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 194–201 (cit. on p. 15).

Durante, Daniele and David B. Dunson (2014). “Nonparametric Bayes dynamic modelling of relational data”. *Biometrika* 101.4, pp. 883–898 (cit. on p. 16).

Durante, Daniele and David B. Dunson (2016). “Locally adaptive dynamic networks”. *Ann. Appl. Stat.* 10.4, pp. 2203–2232 (cit. on p. 34).

Ferguson, Thomas S. (1973). “A Bayesian Analysis of Some Nonparametric Problems”. *The Annals of Statistics* 1.2, pp. 209–230 (cit. on pp. 16, 49).

Foti, Nicholas et al. (2013). “A unifying representation for a class of dependent random measures”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 20–28 (cit. on pp. 9, 16).

Foulds, James et al. (2013). “Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 446–454 (cit. on p. 75).

Foulds, James R. et al. (2011). “A Dynamic Relational Infinite Feature Model for Longitudinal Networks”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 287–295 (cit. on pp. 1, 12, 15–17, 30, 41, 64).

Fox, Emily B. et al. (2011). “A sticky HDP-HMM with application to speaker diarization”. *Ann. Appl. Stat.* 5, pp. 1020–1056 (cit. on p. 75).

Fu, Wenjie et al. (2009). “Dynamic mixed membership blockmodel for evolving networks”. In: *International Conference on Machine Learning*, pp. 329–336 (cit. on pp. 12, 16).

Fujiwara, Yusuke, Yoichi Miyawaki, and Yukiyasu Kamitani (2009). “Estimating image bases for visual image reconstruction from human rain activity”. In: *Advances in Neural Information Processing Systems*, pp. 576–584 (cit. on p. 91).

Gershman, Samuel et al. (2015). “Distance Dependent Infinite Latent Feature Models”. *IEEE Trans. Pattern Anal. Mach. Intell.* 37.2, pp. 334–345 (cit. on p. 24).

Ghahramani, Zoubin et al. (1996). “Factorial Hidden Markov Models”. In: *Advances in Neural Information Processing Systems*, pp. 472–478 (cit. on pp. 1, 12).

Ghosn, Faten et al. (2004). “The MID3 Data Set, 1993-2001: Procedures, Coding Rules, and Description”. *Conflict Management and Peace Science* 21.2, pp. 133–154 (cit. on pp. 15, 41).

Goldenberg, Anna et al. (2010). “A Survey of Statistical Network Models”. *Found. Trends Mach. Learn.* 2.2, pp. 129–233 (cit. on p. 11).

Gopalan, Prem et al. (2015). “Scalable Recommendation with Hierarchical Poisson Factorization”. In: *Annual Conference on Uncertainty in Artificial Intelligence*, pp. 326–335 (cit. on p. 16).

Gopalan, Prem K et al. (2014). “Content-based recommendations with Poisson factorization”. In: *Advances in Neural Information Processing Systems 27*, pp. 3176–3184 (cit. on p. 13).

Guhaniyogi, Rajarshi et al. (2014). “Bayesian Conditional Density Filtering for Big Data”. *CoRR* abs/1401.3632 (cit. on p. 23).

Guo, Fan et al. (2007). “Recovering temporally rewiring networks: a model-based approach”. In: *International Conference on Machine Learning*, pp. 321–328 (cit. on pp. 1, 16).

Gupta, Sunil Kumar et al. (2012a). “A Bayesian Nonparametric Joint Factor Model for Learning Shared and Individual Subspaces from Multiple Data Sources”. In: *SIAM conference on Data Mining*, pp. 200–211 (cit. on p. 75).

Gupta, Sunil Kumar et al. (2012b). “A Slice Sampler for Restricted Hierarchical Beta Process with Applications to Shared Subspace Learning”. In: *Annual Conference on Uncertainty in Artificial Intelligence*, pp. 316–325 (cit. on p. 75).

Hamilton, William L. et al. (2017). “Representation Learning on Graphs: Methods and Applications”. *IEEE Data Eng. Bull.* 40.3, pp. 52–74 (cit. on p. 95).

Hawkes, Alan G. (1971). “Spectra of some self-exciting and mutually exciting point processes”. *Biometrika* 58.1, pp. 83–90 (cit. on pp. 12, 47).

Heaukulani, Creighton et al. (2013). “Dynamic Probabilistic Models for Latent Feature Propagation in Social Networks”. In: *International Conference on Machine Learning*, pp. 275–283 (cit. on pp. 1, 12, 15–17, 41, 64).

Hill, Steven M. et al. (2012). “Bayesian Inference of Signaling Network Topology in a Cancer Cell Line”. *Bioinformatics* 28.21, pp. 2804–2810 (cit. on p. 95).

Ho, Qirong et al. (2011). “Evolving Cluster Mixed-Membership Blockmodel for Time-Evolving Networks”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 342–350 (cit. on pp. 12, 16).

Hoef, Jay M. Ver (2012). “Who Invented the Delta Method?” *The American Statistician* 66.2, pp. 124–127 (cit. on p. 79).

Hoff, P.D. (2008). “Modeling homophily and stochastic equivalence in symmetric relational data”. In: *Advances in Neural Information Processing Systems*, pp. 657–664 (cit. on p. 25).

Hoff, Peter D. et al. (2001). “Latent Space Approaches to Social Network Analysis”. *Journal of the American Statistical Association* 97, pp. 1090–1098 (cit. on pp. 11, 16, 57).

Holland, Paul W. et al. (1983). “Stochastic blockmodels: First steps” (cit. on p. 11).

Hu, Changwei et al. (2015). “Zero-truncated Poisson Tensor Factorization for Massive Binary Tensors”. In: *Annual Conference on Uncertainty in Artificial Intelligence*, pp. 375–384 (cit. on pp. 15, 23).

Hu, Changwei et al. (2016a). “Non-negative Matrix Factorization for Discrete Data with Hierarchical Side-Information”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 1124–1132 (cit. on p. 13).

Hu, Changwei et al. (2016b). “Topic-Based Embeddings for Learning from Large Knowledge Graphs”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1133–1141 (cit. on p. 23).

Ishiguro, Katsuhiko et al. (2010). “Dynamic Infinite Relational Model for Time-varying Relational Data”. In: *Advances in Neural Information Processing Systems*, pp. 919–927 (cit. on p. 16).

Ishiguro, Katsuhiko et al. (2017). “Averaged Collapsed Variational Bayes Inference”. *Journal of Machine Learning Research* 18.1, pp. 1–29 (cit. on p. 76).

Junuthula, Ruthwik R. et al. (2017). “The Block Point Process Model for Continuous-Time Event-Based Dynamic Networks”. *CoRR* abs/1711.10967 (cit. on p. 47).

Kapoor, J., A. Vergari, M. Gomez Rodriguez, and I. Valera (2018). “Bayesian Nonparametric Hawkes Processes”. In: *Bayesian Nonparametrics workshop at the 32nd Conference on Neural Information Processing Systems* (cit. on p. 52).

Kemp, Charles et al. (2006). “Learning Systems of Concepts with an Infinite Relational Model”. In: *AAAI Conference on Artificial Intelligence*, pp. 381–388 (cit. on p. 11).

Kim, Dae Il et al. (2012). “The Nonparametric Metadata Dependent Relational Model”. In: *International Conference on Machine Learning*, pp. 1411–1418 (cit. on pp. 13, 95).

Kim, Myunghwan et al. (2013). “Nonparametric Multi-group Membership Model for Dynamic Networks”. In: *Advances in Neural Information Processing Systems*, pp. 1385–1393 (cit. on pp. 12, 16, 17, 24, 41, 64).

Kim, Myunghwan and Jure Leskovec (2012). “Latent Multi-group Membership Graph Model”. In: *International Conference on Machine Learning*, pp. 947–954 (cit. on p. 95).

Kingman, J. F. C. (1967). “Completely random measures.” *Pacific J. Math.* 21.1, pp. 59–78 (cit. on p. 8).

Kingman, J. F. C. (1993). *Poisson processes*. New York: OUP (cit. on p. 9).

Klami, Arto, Seppo Virtanen, and Samuel Kaski (2013). “Bayesian Canonical Correlation Analysis”. *Journal of Machine Learning Research* 14.1, pp. 965–1003 (cit. on p. 91).

Knowles, D A and Z Ghahramani (2011). “Nonparametric Bayesian sparse factor models with application to gene expression modeling”. *Ann. Appl. Stat.* 5.2B, pp. 1534–1552 (cit. on pp. 75, 76).

Lahti, Leo et al. (2013). “Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review”. *Briefings in Bioinformatics* 14.1, pp. 27–35 (cit. on p. 91).

Lewis, Erik A. and George O. Mohler (2011). “A Nonparametric EM algorithm for Multiscale Hawkes Processes”. *Journal of Nonparametric Statistics* (cit. on p. 54).

Li, Chunyuan et al. (2016). “Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks”. In: *AAAI Conference on Artificial Intelligence*, pp. 1788–1794 (cit. on p. 69).

Linderman, Scott and Ryan Adams (2014). “Discovering Latent Network Structure in Point Process Data”. In: *International Conference on Machine Learning*, pp. 1413–1421 (cit. on p. 47).

Ma, Yi-An et al. (2015). “A Complete Recipe for Stochastic Gradient MCMC”. In: *Advances in Neural Information Processing Systems*, pp. 2917–2925 (cit. on pp. 63, 64, 72).

Mastrandrea, Rossana et al. (2015). “Contact patterns in a high school: A comparison between data collected using wearable sensors”. *PLoS ONE* 10.9, pp. 1–26 (cit. on pp. 38, 71).

Mavroforakis, Charalampos, Isabel Valera, and Manuel Gomez-Rodriguez (2017). “Modeling the Dynamics of Online Learning Activity”. In: *International World Wide Web Conference*, pp. 1421–1430 (cit. on pp. 47, 49, 52).

Miller, Kurt et al. (2009). “Nonparametric Latent Feature Models for Link Prediction”. In: *Advances in Neural Information Processing Systems*, pp. 1276–1284 (cit. on pp. 1, 11, 16).

Miscouridou, Xenia et al. (2018). “Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data”. In: *Advances in Neural Information Processing Systems*, pp. 2343–2352 (cit. on pp. 10, 12, 13, 47–50, 52).

Miyawaki, Yoichi et al. (2008). “Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders”. *Neuron* 60, pp. 915–929 (cit. on p. 92).

Mucha, Peter J et al. (2010). “Community structure in time-dependent, multiscale, and multiplex networks”. *Science* 328.5980, pp. 876–878 (cit. on p. 15).

Newman, M. E. J. and M. Girvan (2004). “Finding and evaluating community structure in networks”. *Phys. Rev. E* 69, p. 026113 (cit. on p. 11).

Nickel, M. et al. (2016). “A Review of Relational Machine Learning for Knowledge Graphs”. *Proceedings of the IEEE* 104.1, pp. 11–33 (cit. on p. 102).

Nowicki, Krzysztof and Tom A. B Snijders (2001). “Estimation and Prediction for Stochastic Blockstructures”. *Journal of the American Statistical Association* 96.455, pp. 1077–1087 (cit. on p. 11).

Oates, Chris J. et al. (2014). “Causal network inference using biochemical kinetics”. *Bioinformatics* 30.17, pp. 468–474 (cit. on p. 95).

Paisley, John and Lawrence Carin (2009). “Nonparametric Factor Analysis with Beta Process Priors”. In: *International Conference on Machine Learning*. New York, pp. 777–784 (cit. on pp. 75, 76).

Palla, Gergely et al. (2005). “Uncovering the overlapping community structure of complex networks in nature and society”. *Nature* 435, pp. 814–818 (cit. on p. 11).

Patterson, Sam et al. (2013). “Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex”. In: *Advances in Neural Information Processing Systems*, pp. 3102–3110 (cit. on pp. 63, 67, 69, 72).

Phan, Tuan Q. and Edoardo M. Airoldi (2015). “A natural experiment of social network formation and dynamics”. *Proceedings of the National Academy of Sciences* 112.21, pp. 6595–6600 (cit. on p. 15).

Pitman, J. (2006). *Combinatorial stochastic processes*. Lectures on Probability Theory. Berlin: Springer-Verlag, pp. x+256 (cit. on pp. 8, 79).

Polson, Nicholas G. et al. (2013). “Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables”. *Journal of the American Statistical Association* 108.504, pp. 1339–1349 (cit. on pp. 15, 26, 28, 54, 73).

Rai, Piyush et al. (2015). “Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings”. In: *Advances in Neural Information Processing Systems*, pp. 3222–3230 (cit. on p. 52).

Rai, Piyush (2017). “Non-negative Inductive Matrix Completion for Discrete Dyadic Data”. In: *AAAI Conference on Artificial Intelligence*, pp. 2499–2505 (cit. on pp. 13, 95).

Rai, Piyush and Hal Daume III (2008). “The Infinite Hierarchical Factor Regression Model”. In: *Advances in Neural Information Processing Systems*. Vancouver, Canada, pp. 1321–1328 (cit. on pp. 75, 76).

Ranganath, Rajesh et al. (2015). “Deep Exponential Families”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 762–771 (cit. on p. 16).

Sarkar, Purnamrita et al. (2006). “Dynamic Social Network Analysis using Latent Space Models”. In: *Advances in Neural Information Processing Systems*, pp. 1145–1152 (cit. on pp. 1, 12, 16).

Schein, Aaron et al. (2016a). “Bayesian Poisson Tucker Decomposition for Learning the Structure of International Relations”. In: *International Conference on Machine Learning*, pp. 2810–2819 (cit. on p. 47).

Schein, Aaron et al. (2016b). “Poisson-Gamma dynamical systems”. In: *Advances in Neural Information Processing Systems*, pp. 5005–5013 (cit. on p. 16).

Schein, Aaron et al. (2018). “Locally Private Bayesian Inference for Count Models”. *CoRR* abs/1803.08471 (cit. on p. 102).

Scott, James G. and Liang Sun (2013). “Expectation-Maximization for logistic regression”. *arXiv preprintarXiv:1306.0040* (cit. on p. 55).

Tan, Xi et al. (2018). “Nested CRP with Hawkes-Gaussian Processes”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1289–1298 (cit. on pp. 13, 49, 57).

Tan, Xi, Vinayak Rao, and Jennifer Neville (2018). “The Indian Buffet Hawkes Process to Model Evolving Latent Influences”. In: *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 795–804 (cit. on pp. 47–49, 57).

Teh, Y. W. et al. (2006). “A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation”. In: *Advances in Neural Information Processing Systems*, pp. 1353–1360 (cit. on p. 75).

Teh, Y. W. et al. (2007). “Hierarchical Dirichlet Processes”. *Journal of the American Statistical Association* 101.476, pp. 1566–1581 (cit. on pp. 8, 49, 75, 80).

Teh, Y. W. et al. (2008). “Collapsed Variational Inference for HDP”. In: *Advances in Neural Information Processing Systems*, pp. 1481–1488 (cit. on pp. 75, 77).

Thibaux, R. et al. (2007). “Hierarchical beta processes and the Indian buffet process”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 564–571 (cit. on pp. 75, 76).

Todeschini, Adrien et al. (2017). “Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities”. *ArXiv e-prints*. arXiv: `arXiv:1602.02114` (cit. on p. 47).

Virtanen, Seppo, Arto Klami, Suleiman A. Khan, and Samuel Kaski (2012). “Bayesian Group Factor Analysis”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1269–1277 (cit. on pp. 75, 86).

Wainwright, Martin J. and Michael I. Jordan (2008). “Graphical Models, Exponential Families, and Variational Inference”. *Found. Trends Mach. Learn.* (Cit. on pp. 75, 83).

Wang, Pengyu et al. (2013). “Collapsed Variational Bayesian Inference for Hidden Markov Models”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 599–607 (cit. on p. 75).

West, Mike (2003). “Bayesian Factor Regression Models in the "Large p, Small n" Paradigm”. In: *Bayesian Statistics*, pp. 723–732 (cit. on p. 75).

Witten, Daniela M., Trevor Hastie, and Robert Tibshirani (2009). “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis”. *Biostatistics* 10.3, pp. 515–534 (cit. on p. 75).

Wolpert, Robert L. et al. (1998). “Poisson/gamma random field models for spatial statistics”. *Biometrika* 85.2, pp. 251–267 (cit. on p. 9).

Wolpert, Robert L., Merlise A. Clyde, and Chong Tu (2011). “Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels”. *Ann. Statist.* 39.4, pp. 1916–1962 (cit. on p. 9).

Xing, Eric P. et al. (2010). “A state-space mixed membership blockmodel for dynamic network tomography”. *Ann. Appl. Stat.* 4.2, pp. 535–566 (cit. on pp. 1, 12, 16).

Xu, Hongteng et al. (2016a). “Learning Granger Causality for Hawkes Processes”. In: *International Conference on Machine Learning*, pp. 1717–1726 (cit. on pp. 12, 47).

Xu, Hongteng et al. (2016b). “Learning Granger Causality for Hawkes Processes”. In: *International Conference on Machine Learning*, pp. 1717–1726 (cit. on p. 54).

Xu, Hongteng and Hongyuan Zha (2017). “A Dirichlet Mixture Model of Hawkes Processes for Event Sequence Clustering”. In: *Advances in Neural Information Processing Systems 30*, pp. 1354–1363 (cit. on p. 47).

Xu, Kevin (2015). “Stochastic Block Transition Models for Dynamic Networks”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087 (cit. on p. 24).

Xu, Kevin S. et al. (2014). “Dynamic Stochastic Blockmodels for Time-Evolving Social Networks”. *J. Sel. Topics Signal Processing* 8.4, pp. 552–562 (cit. on pp. 12, 16, 39, 41, 70).

Yang, Jiasen et al. (2017). “Decoupling Homophily and Reciprocity with Latent Space Network Models”. In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence* (cit. on pp. 10, 12, 13, 47, 57).

Yang, S. and H. Koeppl (2018a). “Collapsed Variational Inference for Nonparametric Bayesian Group Factor Analysis”. In: *IEEE International Conference on Data Mining (ICDM)*, pp. 687–696 (cit. on p. 4).

Yang, Sikun and Heinz Koeppl (2018b). “A Poisson Gamma Probabilistic Model for Latent Node-Group Memberships in Dynamic Networks”. In: *AAAI Conference on Artificial Intelligence*, pp. 4366–4373 (cit. on pp. 4, 70).

Yang, Sikun and Heinz Koeppl (2018c). “Dependent Relational Gamma Process Models for Longitudinal Networks”. In: *International Conference on Machine Learning (ICML)*, pp. 5551–5560 (cit. on p. 4).

Yang, Sikun and Heinz Koeppl (2019a). “An Empirical Study of Stochastic Gradient MCMC Algorithms for the Dynamic Edge Partition Models”. In: *Submitted* (cit. on p. 4).

Yang, Sikun and Heinz Koeppl (2019b). “The Hawkes Edge Partition Model for Continuous-time Event-based Temporal Networks”. In: *Submitted* (cit. on p. 4).

Zhang, Boqian et al. (2017). “Collapsed variational Bayes for Markov jump processes”. In: *Advances in Neural Information Processing Systems*, pp. 3749–3757 (cit. on p. 76).

Zhang, Quan and Mingyuan Zhou (2018). “Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks”. In: *Advances in Neural Information Processing Systems*, pp. 5002–5013 (cit. on p. 52).

Zhao, Shiwen et al. (2016). “Bayesian group factor analysis with structured sparsity”. *Journal of Machine Learning Research* 17.196, pp. 1–47 (cit. on pp. 86, 87).

Zhou, Ke, Hongyuan Zha, and Le Song (2013). “Learning Triggering Kernels for Multi-dimensional Hawkes Processes”. In: *International Conference on Machine Learning*, pp. 1301–1309 (cit. on p. 54).

Zhou, Mingyuan et al. (2012). “Lognormal and Gamma Mixed Negative Binomial Regression.” In: *International Conference on Machine Learning*, pp. 1343–1350 (cit. on p. 54).

Zhou, Mingyuan et al. (2015). “Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1135–1143 (cit. on pp. 15, 16, 18, 24, 25, 30, 47, 49, 50, 52, 57, 65, 95).

Zhou, Mingyuan et al. (2016). “Augmentable Gamma Belief Networks”. *Journal of Machine Learning Research* 17, pp. 1–44 (cit. on pp. 16, 73).

Zhou, Mingyuan (2016). “Softplus Regressions and Convex Polytopes”. *arXiv:1608.06383* (cit. on p. 55).

Zhou, Mingyuan (2018a). “Discussion on “Sparse graphs using exchangeable random measure” by Francois Caron and Emily B. Fox”. *arXiv preprint arXiv:1802.07721* (cit. on p. 50).

Zhou, Mingyuan (2018b). “Parsimonious Bayesian deep networks”. In: *Advances in Neural Information Processing Systems*, pp. 3190–3200 (cit. on p. 52).

Zhou, Mingyuan and Lawrance Carin (2015a). “Negative Binomial Process Count and Mixture Modeling”. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 37.2, pp. 307–320 (cit. on pp. 8, 48, 49, 52).

Zhou, Mingyuan and Lawrence Carin (2015b). “Negative Binomial Process Count and Mixture Modeling”. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 37.2, pp. 307–320 (cit. on pp. 15, 16, 26, 65, 66, 75).

Zitnik, M. et al. (2019). “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. *Information Fusion* 50, pp. 71–91 (cit. on p. 95).

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2006). “Sparse Principal Component Analysis”. *J. Comput. and Graph. Statist.* 15.2, pp. 265–286 (cit. on p. 75).

CURRICULUM VITÆ

SIKUN YANG

PERSONAL INFORMATION

DATE OF BIRTH	24 June, 1986
PLACE OF BIRTH	Jilin, P.R. China

EDUCATION

Master of Philosophy (M.Phil.) in Computer Science and Technology September 2011-June 2014
Beijing Jiaotong University, Beijing, P.R. China

Bachelor of Science (B.Sc.) in Electrical Engineering September 2005-June 2009
Beijing Jiaotong University, Beijing, P.R. China

WORK EXPERIENCE

Technische Universität Darmstadt May 2014-August 2019
Research Associate, Bioinspired Communication Systems
Darmstadt, Germany

January 23, 2020

ERKLÄRUNG LAUT §9 PROMOTIONSORDNUNG

§ 8 Abs. 1 lit. c PromO
Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§ 8 Abs. 1 lit. d PromO
Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§ 9 Abs. 1 PromO
Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§ 9 Abs. 2 PromO
Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Datum und Unterschrift