Faculty of Engineering and Information Technology
University of Technology, Sydney

Nonparametric Bayesian Models for Learning Network Coupling Relationships

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Xuhui Fan

May 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

To my parents and my wife

for your love and support

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Longbing Cao for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I also would like to appreciate my co-supervisor Dr. Richard Yi Da Xu, for leading into the wonderful world of nonparametric Bayesian methods and for the sleepless nights we were working together before the deadlines. Without his professional guidance and persistent help, this thesis would not have been possible.

I thank my fellow labmates in Advanced Analystics Institute: Lin Zhu, Xiaodong Yue, Yin Song, Wei Wei, Can Wang, Zhong She, etc., for our stimulating discussions and for all the fun we have had in the last four years.

Also, I would like to thank the magic of machine learning. This fantastic world has made me lucky enough to have the fun while exploring it during all of my Ph.D period.

Last but not the least, I would like to thank my family: my wife, my parents and my parents in law, for their unconditional support, both financially and emotionally throughout the whole PhD studying.

Xuhui Fan
December 2014 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Papers Published

- **Xuhui Fan**, Longbing Cao, Richard Yi Da Xu. Dynamic Infinite Mixed-Membership Stochastic Blockmodel. Accepted in IEEE Transaction on Neural Networks and Learning Systems.
- **Xuhui Fan** and Longbing Cao. A Convergence Theorem for the Graph Shift-type Algorithms. Accept in Pattern Recognition.
- **Xuhui Fan**, Lin Zhu, Longbing Cao, Xia Cui and Yew Soon Ong. Maximum Margin Clustering on Evolutionary Data. Proceedings of the 21th ACM international conference on Information and knowledge management(ACM CIKM). October 2012.
- Wei Wei, **Xuhui Fan**, Jinyan Li and Longbing Cao. Model the Complex Dependence Structures of Financial Variables by Using Canonical Vine. Proceedings of the 21th ACM international Conference on Information and knowledge management(ACM CIKM). October 2012.
- **Xuhui Fan** and Longbing Cao. A theoretical framework of the Graph Shift Algorithm. Proceedings of the Twenty-Sixth AAAI Conference on Arti.cial Intelligence(AAAI-12 poster). July 2012.

Papers to be Submitted/Under Review

- **Xuhui Fan**, Eric Gaussier, Longbing Cao, Richard Yi Da Xu. Supervised Categorical Metric Learning with Schattern p-norm. Submitted

to SDM 2015.

- **Xuhui Fan**, Richard Yi Da Xu, Longbing Cao, Yin Song. Learning Hidden Structures with Relational Models by Adequately Involving Rich Information in A Network. CoRR abs/1310.1545(2013) Submitted to AISTATS 2015.

- **Xuhui Fan**, Longbing Cao, Richard Yi Da Xu. Copula Mixed-Membership Stochastic Blockmodel for Intra-Subgroup Correlations. CoRR abs/1306.2733(2013) Submitted to IEEE Transaction on Pattern Analysis and Machine Intelligence.

- **Xuhui Fan**, Yiling Zeng, Longbing Cao. Non-parametric Power-law Data Clustering. CoRR abs/1306.3003(2013).

Abstract

As the traditional machine learning setting assumes that the data are identically and independently distributed (i.i.d), this is quite like a perfect conditioned vacuum and seldom a real case in practical applications. Thus, the non-i.i.d learning (Cao, Ou, Yu & Wei 2010)(Cao, Ou & Yu 2012)(Cao 2014) has emerged as a powerful tool in describing the fundamental phenomena in the real world, as more factors to be well catered in this modelling. One critical factor in the non-i.i.d. learning is the relations among the data, ranging from the feature information, node partitioning to the correlation of the outcome, which is referred to as the coupling relation in the non-i.i.d. learning. In our work, we aim at uncovering this coupling relation with the nonparametric Bayesian relational models, that is, the data points in our work are supposed to be coupled with each other, and it is this coupling relation we are interested in for further investigation. The coupling relation is widely seen and motivated in real world applications, for example, the hidden structure learning in social networks for link prediction and structure understanding, the fraud detection in the transactional stock market, the protein interaction modelling in biology.

In this thesis, we are particularly interested in the learning and inferencing on the relational data, which is to further discover the coupling relation between the corresponding points. For the detail modelling perspective, we have focused on the framework of *mixed-membership stochastic blockmodel*, in which *membership indicator* and *mixed-membership distribution* are noted to represent the nodes' belonging community for one relation and the his-

togram of all the belonging communities for one node. More specifically, we are trying to model the coupling relation through three different aspects: 1) the mixed-membership distributions' coupling relation across the time. In this work, the coupling relation is reflected in the sticky phenomenon between the mixed-membership distributions in two consecutive time; 2) the membership indicators' coupling relation, in which the Copula function is utilized to depict the coupling relation; 3) the node information and mixed-membership distribution's coupling relation. This is achieved by the new proposal transform for the node information's integration. As these three aspects describe the critical parts of the nodes' interaction with the communities, we are hoping the complex hidden structures can thus be well studied.

In all of the above extensions, we set the number of the communities in a nonparametric Bayesian prior (mainly Hierarchical Dirichlet Process), instead of fixing it as in the previous classical models. In such a way, the complexity of our model can grow along with the data size. That is to say, while we have more data, our model can have a larger amount of communities to account for them. This appealing property enables our models to fit the data better. Moreover, the nice formalization of the Hierarchical Dirichlet Process facilitates us to some benefits, such as the conjugate prior. Thus, this nonparametric Bayesian prior has introduced new elements to the coupling relations' learning.

Under this varying backgrounds and scenarios, we have shown our proposed models and frameworks for learning the coupling relations are evidenced to outperform the state-of-the-art methods via literature explanation and empirical results. The outcomes are sequentially accepted by top journals. Therefore, the nonparametric Bayesian models in learning the coupling relations presents high research value and would still be attractive opportunities for further exploration and exploit.

Chapter 1

Introduction

1.1 Problem Statement

There is undoubted evidence that the data analysis and machine learning have had a tremendous growth in recent decades. Its successful applications have been found in areas such as the social network analysis, image processing, speech recognition, and fraud detection under various settings. However, it should be noted that, most of the existing algorithms, models and even theoretical foundations in these related fields, such as machine learning, statistics, and data mining, have made an ideal assumption that data are independently and identically distributed (i.i.d.). This naive assumption helps to simplify the complicated situations we are facing in the real world. Although impressive performances have been seen in some applications, this i.i.d assumption does not hold in many of complex real world problems, such as the social community detection, the stock trading market and recommendation systems (Cao et al. 2010)(Cao et al. 2012)(Cao 2014) . In these real-life cases, data points have certain relationships such that learning these relationships would be a necessary task to promote a better understanding to them. Thus, there is an emergent need to break this i.i.d. assumption in analysing the data. That is to say, we need to do the modelling from the non-i.i.d. perspective.

In general, the non-i.i.d. learning covers many important fields, including the coupling relation learning and heterogeneous data modelling. In this thesis, there is a focus on the study of the objects' coupling relations, which is far more meaningful and complicated than the usual dependency relation. In real world, the coupling relation can be represented in various forms on different objects. For instance, one behaviour happens after another (serial coupling), one behaviour causes the occurrence of another (casual coupling), two behaviours happen at the same time due to the same reason (synchronous coupling), different events happen on a mutually exclusive basis (exclusive coupling) and some behaviors or social events have required dependents such as prefix or postfix components (dependent coupling) (Cao 2014). Also, the coupling relation occurs in different levels of the data, ranging from the level of data attributes value to the level of the whole group construction. All these real and complicated scenarios require the consideration of the above coupling relations while trying to model and learn from the data.

The network structure analysis, for instance, is a typical example where analyzing the coupling relationships become necessary. The classical network analysis algorithms (from the Bayesian point) usually involve the element of the similarity (or distance) matrix, which is often measured within the nodes. This has greatly correlated to the relations between the nodes. Based on these descriptions on correlations (which we are referred to as coupling relations), it is hoped to use network analysis to handle the issues such as: (1) a better representation of networks, which should be able to integrate different sources of information in a network; (2) linkage prediction and identification, which is to fully understand the linkages within the network; (3) discovery of key nodes, which may represent the centers of the network; (4) the evolvement of the network with time.

However, as has been found in the literature review, little concern has been given to the understanding the above issues from the coupling relations point of view. These coupling relations are closely related in various levels of the network, including the communities inside the network, the net-

work's nodes and feature information of the nodes. Correspondingly, the coupling relations are presented within and between these levels, such as the coupling relation on the dynamic behaviour of communities, the coupling relation within the communities themselves, the coupling relation on the communities' distribution between the nodes and the coupling relation on nodes' attribute information to communities. As can be seen, these coupling relations often present several critical properties among the whole network, which is usually ignored in the classical models.

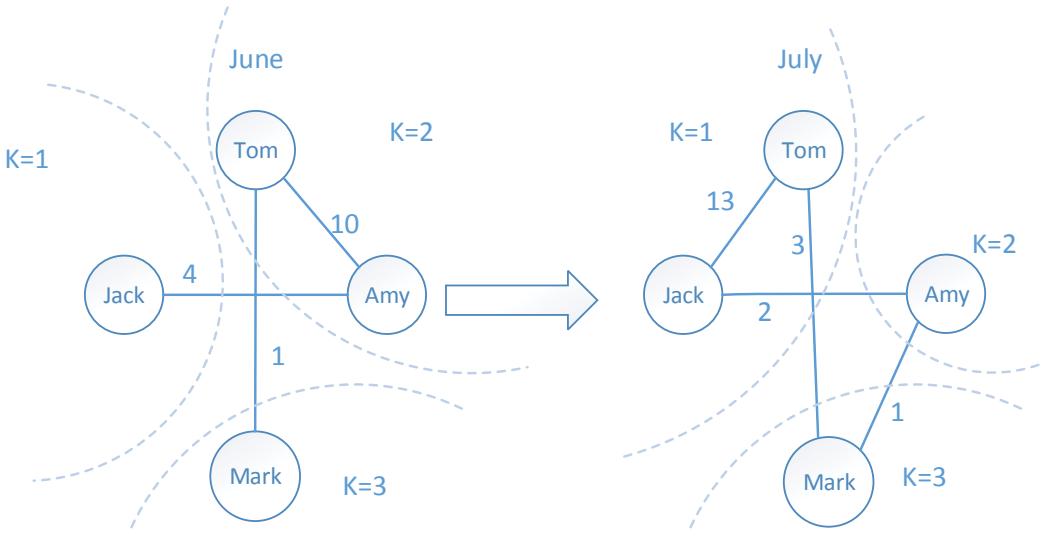

Figure 1.1: The network structure.

Table 1.1: Persons' Information

Person	age	occupation	salary	marriage	hobby	education
Tom	35	salesman	middle	yes	fishing	high school
Amy	37	manager	high	yes	fishing	university
Jack	43	doctor	high	yes	reading	university
Mark	25	student	low	no	football	phd

To illustrate these issues more clearly, an example is used, in which Figure

1.1 and Table 1.1 display four persons' social activities during the two consecutive months (that is, June and July) and their corresponding profiles, which include the attribute information of age, occupation, salary, marriage, hobby and education. In Figure 1.1, the number next to the link represents the two persons' communication times in each month and K denotes the community index. As can be seen, these four persons' activities change dramatically across time. From June to July, Tom and Amy reduced their communication times from 10 to zero, while Tom and Jack started their communications from none to 13 times in July. Correspondingly, their community memberships change as Tom transfers from community 2 (that is, the same community with Amy) to community 1 (that is, the same community with Jack). In the profile information of Table 1.1, these four people would share both similarities and dissimilarities. For instance, Tom and Amy have the same hobby and similar ages; however, their salaries and educations are different.

This simple example leaves many crucial questions, some of which are listed here:

- Why would Tom stop communicating with Amy? Is it due to the changes in their communities' memberships or the increased communications with Jack or his profile information?
- Is this community partition reasonable to characterize the network structure? Why does Tom and Mark still belong to different communities even if they have stable communications?
- How can the communities' number be known in advance?
- How does the node profile information affect the nodes' belonging communities? Can the most relevant information to this be found?

From the coupling relations perspective, these questions individually correspond to the coupling relations between the nodes in the network, between the communities within the network, within the whole communities' performance and between the ones profile information and communities. It is

believed that a proper study in the coupling relations of this network would no doubt promote a better understanding of the networks.

As shown above, even in this network of merely four people, many challenges and questions have been raised. In reality, a social network even at a moderate size may involve hundreds or thousands of people. These issues are no doubt growing exponentially. Here are summarized several critical issues which include:

- **A proper representation on the data structure**. As shown in the above example, the available information can come in different forms and there are various concepts needed in the modelling. Thus, a proper description on the data structure is needed, as well as using these descriptions to construct the models.

- **Coupling relations between objects over time.** In the above example, Tom behaves significantly in consecutive times. Apart from this temporal observations' change, can the node's membership change be inferred? Moreover, can the compatibilities' change among the hidden communities be identified?

- **Coupling relations within the communities of networks.** In the above example, the four persons belong to 3 different communities. Can the ways of these different communities' influence on the relation within themselves be identified? That is to say, can the intra-communities' coupling relation be efficiently modelled?

- **Coupling relations between the node information and their belonging communities.** As all of the four persons Tom, Amy, Jack and Mark possess sufficient personal information, can it be identified if this information would promote the network partition? If so, which of it is more important among them?

- **A proper inference on these coupling relations.** After the modelling of these coupling relations, their efficient and effective inference

remains a critical issue. Moreover, as this study would like to consider the coupling relations inside the network modelling, their efficient inference with no or less additional computational cost is a desired property.

1.2 Research Methods

1.2.1 Relational Models

The relational models can be powerful tools to deal with the challenges mentioned above. In general, the relational models aim at the understanding of the hidden network structure by partitioning the network into several communities, and these partitions are usually based on the observed relational data. Among many of its proposed interesting settings, the "mixed-membership distribution" (that is, the histogram of the node's belonging communities in generating the observed relational data) plays a critical role in the modelling and has facilitated the expressive description of the whole network structure, as well as the complicated coupling relations among it.

When looking at it more closely, the mixed-membership distribution bridges between the communities, the node and even the node's attribute information. For instance, the node would belong to one or several communities; this belonging relation is heavily dependent on its attribute information. Other aspects also cover several aspects of the network structure, including the hidden partitioning communities, the communities' behaviour of each node, the communities' compatibilities. Based on these practical concepts, the coupling relation can be carefully described.

Let us first turn back to a brief description of relational models. In general, the relational models are categorized into two major frameworks: the latent class model (LCM) and the latent feature model (LFM). Both of them assume the observed relation is parameterized by an entry from the community-compatibility matrix (that is, indicating the compatibilities between the communities). This entry is indexed by the two corresponding

nodes' belonging communities. Their main difference is hence in the way this entry is indexed. For LCM, it is assumed that the indices for each pair of nodes are derived from the two associated hidden class labels (that is, index one entry in the role-compatibility matrix), whereas in the case of LFM, it is assumed that the indices are determined from a set of latent features associated with the pair of nodes (i.e. index one row and one column in the role-compatibility matrix).

The mixed-membership distribution has a central role in describing the coupling relation of objects over time. Through its dynamic behaviour, the node's change in its belonging communities can be inferred. Moreover, the compatibilities between the communities can be observed during the time. On the coupling relation within the subgroups of networks, this study focuses on the dependency between the mixed-membership distributions. In this way, the two nodes' relation is not only affected by the compatibilities of communities, but also influenced by the communities' themselves. Further, the mixed-membership distribution can be a bridge to connect the node attribute information with the communities, which is the community-attribute coupling relation. In this case, the attribute information can be incorporated to constitute the components of mixed-membership distribution, which would influence the communities subsequently.

There are two mainstream effective paradigms for the inference of relational models: Markov Chain Monte Carlo inference (MCMC) and variational inference. The idea of variational inference is to approximate the posterior distribution with a simple, tractable proposal distribution by minimizing some criterion, such as the KL-divergence [5]. In MCMC approximation, the idea is to generate a number of random samples from the posterior distribution and approximate intractable integrals and summations by empirical averages based on the samples. In this work, the focus is on the MCMC inference, which is mainly the Gibbs Sampling and Slice Sampling strategy. As the coupling relations among the relational models have been considered here, these enhanced points can be technically integrated to the detail infer-

ence method and thus improve the learning process.

1.2.2 Nonparametric Bayesian Method

The nonparametric Bayesian (NPB) method aims at utilizing the advantages of both the nonparametric and Bayesian fields. In the nonparametric field, the model is preferred to grow its complexity while more data is being observed. The growing complexity is achieved with the development of the stochastic process prior information, which does not need to fix the number of parameters in apriori. The prior information is further learned in the Bayesian field. More specifically, this nonparametric prior would combine with the likelihood to constitute the posterior probability via the Bayes's rule. Then, the corresponding model can be learned via markov chain monte carlo (MCMC) or variational inference.

In the relational model of this thesis, the number of communities or the number of latent features usually measures the model complexity. While these statistics are learned from the data itself, the potential benefits outweigh the convenience of learning. The expressive representation of the NPB prior (especially the Dirichlet Process) helps avoid the risk of over-fitting and under-fitting of the parameters. More importantly, these NPB methods would provide the following benefits of modelling the coupling relations.

- the NPB method makes the representation on the data structure and the model more appropriate. More data means more model complexity presented in this case.

- the NPB method gives prior information on the node's communities' distribution, which is usually the key part of the relational model. Based on this distribution, the coupling relations can be depicted in a powerful way.

- while the Gibbs sampling is being used to do the inference of the relational model, these nonparametric Bayesian methods can be efficiently

learned. More importantly, several technical skills can be used to speed up the inference procedure.

1.3 Aims and objectives

The Coupling Relations occur in various levels of the network and have also been represented in different forms. The aims and objectives of this thesis is to model the following coupling relations:

- **Coupling relations between the mixed-membership distributions across the time**.
- **Coupling relations within the communities of networks**.
- **Coupling relations between the node information and mixed-membership distribution**.

1.4 Research Challenges

Besides the aims listed above, other issues are also highly challenging and interesting, which can be exemplified as:

- **Coupling relations between the communities' compatibility.** The communities' coupling relations have different forms, such as hierarchical, overlapping, exclusive, while they also involve almost all of the network aspects. Thus, an effective and efficient way to fully represent this remains an open question.
- **Coupling relations within the nodes' mixed-membership distribution**. The components of the nodes' mixed-membership distribution would present several correlations. Some tend to take simultaneous effect, while some may be repelled from each other. Thus, this study interested in fully describing these components' interactions.

- **Coupling relations within the nodes' profile information**. There are fruitful structures in the form of the nodes' profile information. An effective way of using these structure would be challenging.

1.4.1 Coupling relations between the mixed-membership distributions across time

The serial coupling relation describes the two objects' relation during the consecutive times. In this relational models setting, the reference is made to the mixed-membership distributions' dynamic behaviour as the serial coupling relation. While the existing work either focuses on the nonparametric extension or static modelling, the cases are considered in a joint manner. Several interesting issues involved in this problem can be listed as follows:

- identifying the most common dynamic behaviours in a network setting;
- deciding which modelling method works best in different scenarios;
- checking out the other statistics except the mixed-membership distribution that is very largely affected by the dynamic environment;
- observing the communities' evolvement during the time.
- comparing various sampling method's differences in inferencing the model

1.4.2 Coupling relations within the communities of networks

The causal coupling and exclusive coupling relations characterize the two objects' positive or negative effects on each other. In this study's relational models setting, this is noted as the communities' influence in generating a single observation, through the mixed-membership distribution. As all of the classical models implicitly assume the membership indicator pair is independently generated, their strong correlations within the same communities are

not well depicted. In more details, the study is trying to solve several issues as listed:

- characterizing the correlations inside the communities
- incorporating the tool of Copula function to describe the correlation
- identifying the most suitable way to describe the coupling relations inside the communities
- employing efficient inference methods in learning the model

1.4.3 Coupling relations between the node information and mixed-membership distribution

The dependent coupling relation is to describe the potential connection between two objects. In the relational models setting, this study is interested in learning this dependent coupling relation between the node attribute information and its mixed-membership distribution. This problem is particularly interesting and challenging as the previous work either was confined to low inefficient inference or inappropriate modelling (Kim, Hughes & Sudderth 2012). The details of problems to be addressed in this study are:

- properly integrating the node attribute information into the mixed-membership distribution
- efficient and effective modelling strategy in the learning process
- the transformation between the parametric and nonparametric cases
- possible extensions to other relational models

1.5 Research Contributions

In addressing the above research issues, the research contributions in this thesis are summarized below.

- Proposed generic models that can deal with the dynamic case and infinite communities case in mixed-membership stochastic blockmodel (chapter 3);

- Provided a comprehensive study on using different sampling methods (that is, Gibbs sampling and slice sampling) in learning the nonparametric Bayesian models (chapter 3);

- Analyzed the convergence behaviours in terms of Gibbs sampling and slice sampling in the dynamic infinite mixed-membership stochastic blockmodel (chapter 3) ;

- Described the casual and exclusive coupling relations within the communities of networks (chapter 4) ;

- Integrated the Copula function into the generation of membership indicator pair in generating the link data to fully utilize the membership indicator pair's correlation (chapter 4) ;

- Using the properties of nonparametric Bayesian methods and Copula function, marginalizing out hidden variables and proposed two collapsed models (chapter 4);

- Described the dependent coupling relations between the node attribute information and its mixed-membership distribution, i.e., incorporated the node attribute information into the latent class model (Chapter 5);

- Developed a conjugate model in incorporating the node-information in to the mixed-membership stochastic blockmodel, and further compared their convergence bahaviour (Chapter 5);

- Naturally extended the nonparametric Bayesian model developed in this study into the finite community case and also the latent feature model (Chapter 5)

1.6 Thesis Structure

The thesis is structured as follows:

Chapter 2 gives background knowledge and related work on several topics specific to this thesis. In here, the nonparametric Bayesian methods are introduced (which includes the Dirichlet Process, the Hierarchical Dirichlet Process, the Indian Buffet Process), the Monte Carlo Markov Chain (MCMC) methods (which includes the Metropolis-Hastings algorithm), as well as the relational models (which focus on the mixed-membership stochastic blockmodel and latent feature relational model). All these introductions serve the purpose of forming a foundation in understanding the core contributions of this thesis. In the related work part, sufficient literature reviews on the relational models (which covers the latent class model, latent feature, dynamic relational model, and other extensions) and the stick-breaking process are given.

Chapter 3 introduces a relational model, which targets extending the current Mixed-Membership Stochastic Blockmodel to the dynamic and infinite state cases. In particular, additional model parameters are introduced to reflect the degree of persistence between one's memberships at consecutive time stamps. Under this framework, two specific models, namely the mixture time variant (MTV) and the mixture time invariant (MTI), are proposed to depict two different time correlation structures, with the first on the mixed-membership distribution and the later on the membership indicator alone. On the inference procedure, both the slice sampling and Gibbs sampling schemes are utilized to learn the model. Their sampling behaviours are deliberately studied through the technique of MCMC analysis.

Chapter 4 tries to incorporate the Copula function in the Mixed-Membership Stochastic Blockmodel to fully exploit each individual node's participation (or membership) in a social network. Despite its powerful representations, MMSB assumes that the membership indicators of each pair of nodes (that is, people) are distributed independently. However, such an assumption often does not hold in real-life social networks, in which certain known groups

of people may correlate with each other in terms of factors such as their membership categories. To expand MMSB's ability to model such dependent relationships, a new framework - Copula Mixed-Membership Stochastic Blockmodel - is introduced in this thesis for modeling intra-group correlations, namely an individual Copula function which jointly models the membership pairs of those nodes within the group of interest. This framework enables various Copula functions to be used on demand, while maintaining the membership indicator's marginal distribution needed for modelling membership indicators with other nodes outside of the group of interest. Sampling algorithms for both the finite and infinite number of groups are also detailed. The experimental results show the frameworks' superior performance in capturing group interactions when compared with the baseline models on both synthetic and real world datasets.

Chapter 5 proposes a conjugate model in integrating the node-information in the modelling of the nodes' mixed-membership distribution. The current existing models utilise only binary directional link data to recover hidden network structures. However, the attribute information associated with each node contains crucial information to help practitioners understand the underlying relationships in a network. For this reason, this study proposes two models and their solutions, namely the node-information involved mixed-membership model (niMM) and the node-information involved latent-feature model (niLF), in an effort to systematically incorporate additional node information. To effectively achieve this aim, node information is used to generate individual sticks of a Stick-Breaking Process. In this way, not only can the need be avoided to pre-specify the number of communities beforehand, the algorithm also encourages that nodes exhibiting similar information have a higher chance of assigning the same community membership. Substantial efforts have been made towards achieving the appropriateness and efficiency of these models, including the use of conjugate priors. The framework and its inference algorithms using real world datasets are evaluated, which shows the generality and effectiveness of the models in capturing implicit network

structures.

Chapter 6 concludes the thesis and outlines the scope for future work.

Figure 1.2 shows the research profile of this thesis.

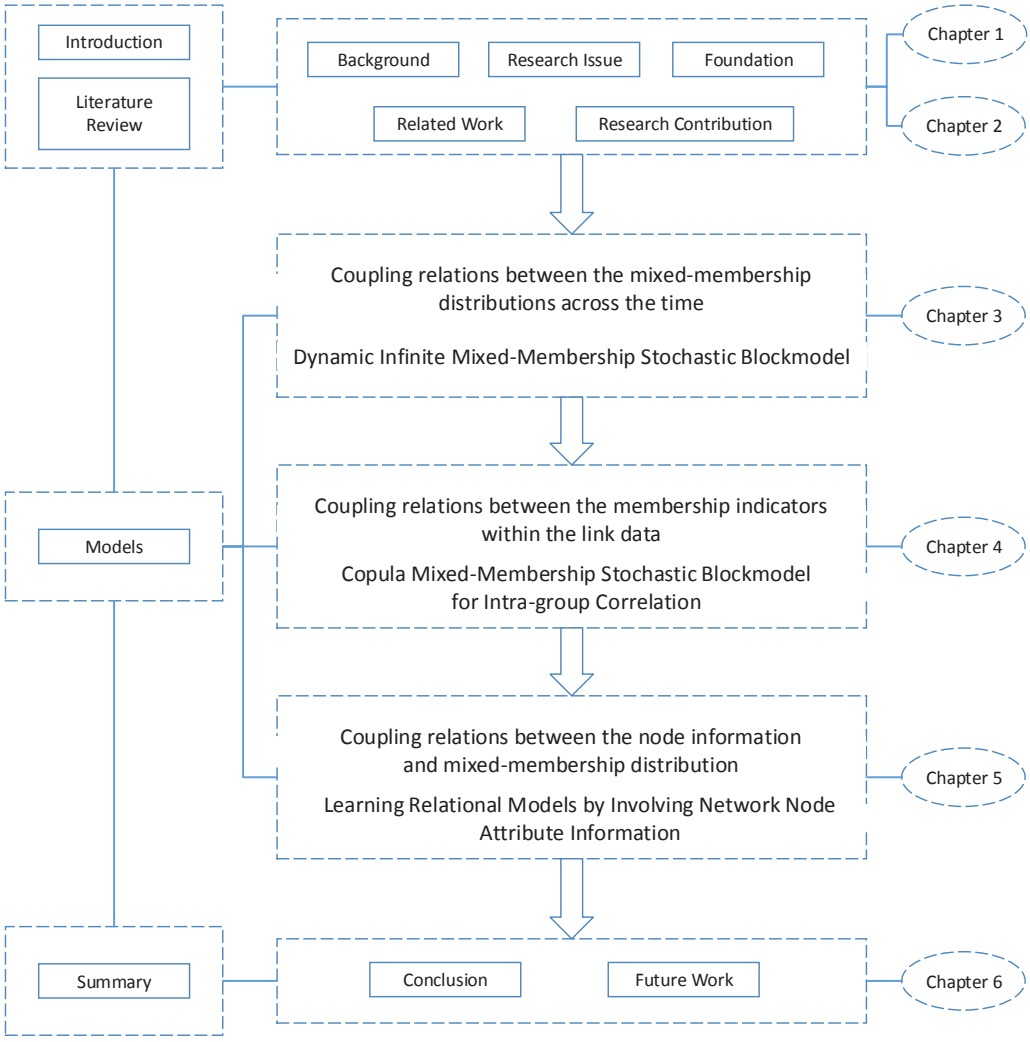

Figure 1.2: The structure of this thesis

Chapter 2

Preliminaries and Literature Review

This chapter provides preliminary knowledge on the nonparametric Bayesian methods, relational models and at the same time, it gives a review on the recent developments and variations of the probabilistic relational models, as well as a literature review on the stick-breaking process, which is commonly used in this thesis. On the introduction to the nonparametric Bayesian methods, the topics to be covered include the Dirichlet Process, the Hierarchical Dirichlet Process, the Beta Process and the Indian Buffet Process. While introducing the relational models, mixed-membership stochastic blockmodel and latent feature relational model are the two detail elaborating models. Their corresponding literature review part comes subsequently, with discussions on their advantages and disadvantages.

2.1 Preliminaries of Nonparametric Bayesian methods

When dealing with a larger size of the dataset in the modelling, it would be natural to expect that the models' parameters would increase correspondingly. That is to say, the parameter set is desired to be adaptive with the

data size and potentially to be infinite when the data size is infinite. To achieve this goal, the nonparametric Bayesian methods are practical solutions by defining prior distributions on function spaces such as the random measures. Thus, the unobserved infinite components' prior can be well defined. Through the classical Monte Carlo Markov Chain (MCMC) inference or the variational inference, these models can be efficiently learned. some classes of nonparametric Bayesian methods are briefly discussed here. For a complete understanding of these methods, refer to (Hjort, Holmes, Müller & Walker 2010).

2.1.1 Dirichlet Process

A Dirichlet process (DP) belongs to a family of stochastic processes whose realizations are distributions, in which these realizations are discrete distributions with countable infinite elements. For each DP, it is uniquely determined by a base measure H and a concentration parameter α, which is to be denoted as $DP(\alpha, H)$. More formally, the Dirichlet process is defined as:

Theorem 2.1 *Let H be a probability distribution on a measurable space Θ, and α a positive scalar. Consider a finite partition $\{A_1, \cdots, A_k\}$ of Θ:*

$$\cup_{k=1}^{K} A_k = \Theta, A_k \cap A_l = \emptyset, \forall k, l \tag{2.1}$$

Based on this definition, we can see that a random probability measure G_0 on Θ is a random draw from a Dirichlet process if its measure on every finite partition follows a Dirichlet distribution:

$$(G_0(A_1), \cdots, G_0(A_k))|\alpha, H \sim Dir(\alpha H(A_1), \cdots, \alpha H(A_K)) \tag{2.2}$$

Theorem 2.2 *For the base measure H and concentration parameter α mentioned above, there exists a unique stochastic process satisfying the above conditions, which we denote by $G_0 \sim DP(\alpha, H)$.*

A practical result upon this definition is that given a set of independent samples generated from G_0, that is, $\theta_1, \cdots, \theta_n \sim G_0$, according to (Ferguson

1973), this gives the posterior distribution of G_0 that:

$$G_0|\theta_1, \cdots, \theta_n, \alpha, H \sim DP(\alpha + n, \frac{\alpha}{\alpha + n}H + \sum_{i=1}^{n} \delta_{\theta_i}) \tag{2.3}$$

where $\delta_\cdot$ is a dirac function.

Stick-breaking process

(Sethuraman 1994) has proved the discreteness of G_0 (that is, G_0 is a discrete distribution) and also provided an explicit construction of G_0:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}, \beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_k), \beta'_k \sim \text{Beta}(1, \alpha), \theta_k \sim H \tag{2.4}$$

This construction process can be set in a stick-breaking analogy. First, a unit stick is broken with ratio β'_1, then on the remaining stick $(1 - \beta'_1)$, the breaking will be done recursively with the ratio of β'_k at the k^{th} step. The generated sticks represents the elements' weight in G_0. As this process can take infinite steps, there would be an infinite number of sticks. This stick-breaking process is also denoted as $\boldsymbol{\beta} \sim GEM(\alpha), \boldsymbol{\beta} = (\beta_1, \cdots, \beta_K, \cdots)$.

As seen from Eq. (2.4), we can see the Dirichlet process generates a discrete distribution with infinite number of components. Moreover, it should be noted that this elegant construction has sparked many interesting variations of the nonparametric Bayesian prior, including the logistic stick-breaking process (Ren, Du, Carin & Dunson 2011), the probit stick-breaking process (Rodriguez & Dunson 2011), the kernel stick-breaking process (Dunson & Park 2008), and the discrete infinite logistic normal process (Paisley, Wang & Blei 2012), all of which aim at replacing the $\beta'_k \sim \text{Beta}(1, \alpha)$ with a new practical form.

Chinese Restaurant Process

The Chinese Restaurant Process (CRP) aims at describing the predictive distribution of a newly appeared variable. Its name is derived from the following analogy: there are customers entering a Restaurant with an infinite

number of tables, and each table is served with a unique dish θ_k. Each arriving customer would choose a table, which is denoted as z_i, with proportioned to the number of customer sitting on this table eating the special dish. With a fixed proportion of α, the customer would choose to sit on a new table $K+1$. Formally speaking, assuming the new customer's dish assignment is $\theta_{\text{new}} = \theta_{z_{n+1}}$ (n customers are already in the restaurant), then this gives

$$\Pr(z_{n+1}|\{z_i\}_{i=1}^n, \alpha) = \frac{\alpha}{\alpha+n}\delta_{K+1} + \frac{1}{n+\alpha}\sum_{k=1}^{K} n_k \delta_k \tag{2.5}$$

Here $n_k = \sum_{i=1}^n \delta(z_i = k)$, denoting the number of customers eating the dish k, $K+1$ denoting the new served dish.

As can be seen, the CRP does not involve the realization of G_0. Actually, the CRP has integrated this G_0 out and uses the posterior distribution to infer the new coming customer's eating dish.

Dirichlet Process Mixture Model

As the Dirichlet process generates random discrete distribution, it is naturally to be used as the weights' prior information in the mixture modelling. While this Dirichlet process prior can help to avoid setting the mixtures' number in advance, it has attracted considerable attentions in the recent decade. It is named the Dirichlet process mixture model, and the simplest form of its generative process is:

$$\theta_k \sim H, \boldsymbol{\beta} \sim \text{GEM}(\alpha), z_i \sim \boldsymbol{\beta}, y_i \sim f(\cdot|\theta_{z_i}), \forall k, i \in \mathbb{N}^+ \tag{2.6}$$

Inferencing the hidden variable $\{z_i\}_{i=1}^n$ in DPMM usually takes a similar form of the Chinese Restaurant Process:

$$\Pr(z_i|z_{\setminus i}, \alpha, \theta, H, f) = \begin{cases} N_k^{-i} \cdot f(y_i|\theta_k); & z_i = k; \\ \alpha \cdot \int_{\theta_{\text{new}}} f(y_i|\theta_{\text{new}}) dH(\theta_{\text{new}}); & z_i = \text{new}. \end{cases} \tag{2.7}$$

Here $N_k^{-i} = N_k - \delta(z_i, k)$, denoting the number of data points belonging to the k^{th} mixture while excluding i.

2.1.2 Hierarchical Dirichlet Process

In many situations, different groups of data are desired to be highly correlated with each other. For instance, in topic modelling, each document is composed of a set of hidden topics and the topic would subsequently generate the words. Moreover, the topics range over different documents and the documents are a mixture of these topic, with different weights. In the mixture modelling, this is to say that different data would have the same mixture, while their mixture weights are different. The Hierarchical Dirichlet Process (Teh, Jordan, Beal & Blei 2006) serves this goal as it takes the base measure G_0 in the Dirichlet Process $DP(\alpha, G_0)$ as a realization from another new Dirichlet Process $DP(\gamma, H)$. More formally, it is represented as:

$$\begin{aligned} G_0 &\sim DP(\gamma, H) \\ G_j &\sim DP(\alpha, G_0), \forall j \in \mathbb{N}^+ \end{aligned} \tag{2.8}$$

The discreteness of G_0 enables the distribution set of $\{G_j\}_{j=1}^n$ to share the same component. Actually, since G_0 has defined fixed normalized weights on these components, $\{G_j\}_{j=1}^n$'s role is to get different normalized weights on these components and their expected mean weights is to be the G_0's. Through this way, not only the components' correlation inside each group can be depicted (through the normalization), but also the components' across groups can also be reflected (they share the same discrete base measure).

Stick-breaking process for HDP

Given the stick-breaking representation for $G_0 = \sum_{i=1}^{\infty} \beta_k \delta_{\theta_k}$, where $\theta_k \sim H, \boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \sim \text{GEM}(\gamma)$, the generated $\{G_j\}_{j=1}^n$ would have different weights on these component $\{\theta_k\}_{k=1}^{\infty}$, which is to be represented as:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_{jk}}, \forall j \in \mathbb{N}^+ \tag{2.9}$$

It should be noted that $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty}$ are independent given the "average" weight $\boldsymbol{\beta}$. Another interesting property is that for any partition of $(\pi_{jk})_{k=1}^{\infty}$,

according to Theorem 2.1, gives:

$$\left(\sum_{k\in\mathcal{K}_1}\pi_{jk},\cdots,\sum_{k\in\mathcal{K}_r}\pi_{jk}\right)\sim \text{Dirichlet}(\sum_{k\in\mathcal{K}_1}\beta_k,\cdots,\sum_{k\in\mathcal{K}_r}\beta_k) \tag{2.10}$$

where $\{\mathcal{K}_l\}_{l=1}^{r}$ is a random partition of $(\pi_{jk})_{k=1}^{\infty}$.

For the detail construction of the $(\pi_{jk})_{k=1}^{\infty}$, it is to be derived as:

$$\begin{aligned}&\beta'_k\sim \text{Beta}(1,\gamma),\beta_k=\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l),\\&\pi'_{jk}\sim \text{Beta}(\alpha\beta_k,\alpha(1-\sum_{l=1}^{k}\beta_l)),\pi_{jk}=\pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl})\end{aligned} \tag{2.11}$$

Due to the discreteness of G_0, the probability of $\{G_j\}_{j=1}^{\infty}$ sharing support points is non-zero. However, if G_0 were chosen to be absolute continuous with respect to the Lebesgue measure, there would be zero probability of the group-specific distributions having overlap support (Fox, Sudderth, Jordan & Willsky 2011*b*).

Chinese Restaurant Franchise

(Teh et al. 2006) has shown that the marginal distribution of Hierarchical Dirichlet process can be described in an analogy as the Chinese Restaurant Franchise, which is an extension to the Chinese Restaurant Process. More specifically, there are J restaurants with an infinite number of the same dishes in each one. This is to denote the group partitions of π_j. Each customer would correspond to one restaurant to enjoy the dish. For the dish allocation in each restaurant, there will be table assignments in the restaurant, and each table serves one dish. The Table distribution here is referred to as the $\{\pi_{jk}\}_{k=1}^{\infty}$ and due to the discreteness of G_0, there is non-zero probability that the two tables in one restaurant would serve the same dish. When a customer enters one restaurant, he would choose the table to each the specific dish, proportional to the number of previous customers sitting on that table, with a fixed proportion to sit on a new table.

There are multiple ways to represent the marginal distribution of the customer's eating dish's alignment. To be notational convenient, the direct-assignment case is chosen, which is to integrate out $G_0, \{G_j\}_{j=1}^J$:

$$p(z_{ji} = k|\boldsymbol{z}^{-ji}, \boldsymbol{m}, \boldsymbol{\beta}) = \begin{cases} (n_{j\cdot k}^{-ji} + \alpha_0\beta_k)f_k^{-ji}(x_{ji}), & k \text{ is previous used;} \\ \alpha_0\beta_u f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}), & k = k^{\text{new}}. \end{cases} \tag{2.12}$$

Hierarchical Dirichlet Process-Hidden Markov Model (HDP-HMM)

A direct application of the Hierarchical Dirichlet Process is the prior setting for the Hidden Markov Model. In a stick-breaking formalism, the generative process of the HDP-HMM can be described as follows (assuming z_0 is randomly sampled):

1. $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$
2. $\boldsymbol{\pi}_k \sim \text{DP}(\alpha_0, \boldsymbol{\beta}), \forall k \in N^+$
3. $\phi_k \sim H, \forall k \in N^+$
4. $z_t \sim \boldsymbol{\pi}_{z_{t-1}}, \forall t = 1, \cdots, T.$
5. $y_t \sim F(\phi_{z_t})$

Here γ, α_0 are the two concentration parameters, $\boldsymbol{\pi}_k$ represents the distribution for the k^{th} component, ϕ_k represents the likelihood parameter, z_t is the latent label for the t^{th} observation and the t^{th} observation y_t is generated from the likelihood distribution $F(\phi_{z_t})$.

As can be obviously seen, the time dependency is reflected in Step 4. The current latent label z_t is determined by the distribution, which is indexed by the previous latent label z_{t-1}. Although this looks quite a simple technique, it has shown its benefits in following sections.

Practical application of this HDP-HMM often lies in the treatment of sequence data, including the time-dynamic data and the speaker recognition data. Also, there are several variants of the HDP-HMM model, including the

sticky-HDP-HMM (Fox, Sudderth, Jordan & Willsky 2008), which models the practical phenomenon that the current latent label would be more likely to be kept in the next time.

2.1.3 Beta Process

The Dirichlet process and its hierarchical extension are described in the previous sections. These processes play an important role in the partition model, which enables the partition number to grow with the increase in the size of the dataset. However, in some cases with there is more concern about to associate a node with binary features to indicate its latent occupied features, instead of the partition requirement. Actually, the partition model can be regarded as the latent feature model with only one binary latent feature being allocated to one data point, while the data point in the Beta Process is characterized by multiple latent features.

In correspondence to the unbounded number of partitions, the Beta process allows for the latent features's number to be potentially infinite. In analogy to the Dirichlet process's Chinese Restaurant process metaphor, the Beta process promotes an Indian Buffet process to promote the sparse subset of features' selection.

Unlike the Dirichlet process' dependent random measure (due to the weights' normalization, actually the Dirichlet process is regarded as normalized Gamma process) basis, the Beta process is set within the framework of Complete Random Measure. That is to say, the samples on disjoint subsets are independent of each other. More formally, consider a probability space Θ, and let B_0 be the base measure on Θ with $B_0(\Theta) = \alpha_0$. While the levy measure is defined (Wang & Carin 2012) on the product space $[0, 1] \times \Theta$:

$$\nu(d\omega, d\theta) = c\omega^{-1}(1-\omega)^{c-1}d\omega B_0(d\theta) \tag{2.13}$$

where $c > 0$ is the concentration parameter in the Beta process and the Beta process is denoted as $BP(c, B_0)$ and a draw $B \sim BP(c, B_0)$ is represented

as:

$$B = \sum_{i=1}^{\infty} \omega_k \delta_{\theta_k} \tag{2.14}$$

Let $q_k \in (0, 1)$ denoting the mass measure of the k^{th} atom, its corresponding weight is sampled as $\omega_k \sim \text{Beta}(cq_k, c(1 - q_k))$.

The Beta-Bernoulli conjugacy promotes a Bernoulli process as:

$$X_i|B \sim BeP(B) \tag{2.15}$$

The detail realization of $\{X_i\}_{i=1}^{\infty}$ is a binary vector, which is often referred to as the latent feature. With special value of $c = 1$, this Beta-Bernoulli process becomes the Indian Buffet Process, which will be elaborated on below.

2.1.4 Indian Buffet Process

As shown by (Teh, Görür & Ghahramani 2007), when $c = 1$, the Beta-Bernoulli Process becomes the Indian Buffet Process, which is to marginalize out B and focus on the predictive task on the latent features.

The Indian Buffet Process analogy can be stated as follows: assuming an Indian Buffet contains an infinite number of dishes, which are referred to as latent features. For the first customer, he would choose the first Possion(α) dishes. Then, the subsequent customers i would choose the previous dish with the rate of $\frac{m_k^{-i}}{i}$ (m_k^{-i} denotes the number of previous customers choosing this dish excluding customer i), and additional new Poisson($\frac{\alpha}{i}$) new dishes.

The above nonparametric Bayesian methods is the foundation to understand the work in this thesis. Since various ideas have been incorporated in each of the following chapters, these ideas would be introduced individually when necessary. The related contents are the hidden Markov model, Copula function and a more detailed introduction to the stick-breaking process.

There is a rich literature in the stick-breaking construction of the Indian Buffet Process (Griffiths & Ghahramani 2005) and its underlying Beta Process (Thibaux & Jordan 2007)(Paisley, Zaas, Woods, Ginsburg & Carin 2010). As have already been seen, the underlying representation under

the Indian Buffet Process is one Beta Process, with the concentration parameter specialized to 1. (Teh et al. 2007) gives us a stick-breaking construction for this specialized beta process as:

$$G = \sum_{k}^{\infty} \pi_k \delta_{\theta_k}, \pi_k = \prod_{l=1}^{k} \psi_l, \psi_k \overset{iid}{\sim} \text{Beta}(\gamma, 1), \theta_k \overset{iid}{\sim} G_0. \tag{2.16}$$

Regarding this special construction for a simplified beta process, a construction of a general Beta Process was proposed by (Paisley et al. 2010), which was later followed by an improved version (Paisley, Blei & Jordan 2012).

2.2 Preliminaries of Markov Chain Monte Carlo (MCMC) methods

As stated in the introduction, the variational inference and the Markov Chain Monte Carlo (MCMC) method are the two mainstream paradigms in inferencing the model with nonparametric Bayesian prior, as the nonparametric Bayesian prior's inference is usually intractable. This section gives an introduction to the MCMC method, since it is the main method used in this thesis. Readers are encouraged to refer to (Müller & Quintana 2004)(Shachter 1998)(Walker, Damien, Laud & Smith 1999) for a complete review on all of the inference methods.

In general, MCMC methods produce a sequence of samples which is able to estimate the desired integration $f(\cdot)$. Its estimation is based on a Monte Carlo integration, which is

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \approx \mathbb{E}_x\left[f(x)\right] = \int_{x\sim\mathcal{P}} f(x)dx \tag{2.17}$$

where $\{x_i\}_{i=1}^{n}$ is the sequence of samples and they are generated from a constructed Markov Chain; $\mathcal{P}$ is x's distribution. The critical point of the MCMC methods is that it is not needed to get an analytical form of $\mathcal{P}$, which is usually impossible to get. Its asymptotic guarantee is provided by the Strong Law of Large Numbers (Gallager 2009).

2.2.1 Metropolis-Hastings Sampling

The Metropolis-Hastings (M-H) algorithm represents a generic form for the constructed markov chain. Its value is depended on the proposal distribution (denoted as $q(\cdot)$ and the posterior likelihood term (denoted as $p(\cdot)$) only. Here an easy-sampled proposal distribution $q(\cdot)$ is used since it would be hard to sample from the true distribution. From an algorithmic point of view, in each step τ, suppose x^* is sampled from proposal distribution $q(x)$ and $x^{\tau-1}$ denotes x's value in $(\tau-1)$ step, then x^* would be accepted as x^τ with the ratio $A_\tau(x^*, x^{\tau-1})$, otherwise take $x^\tau = x^{\tau-1}$. Here the ratio $A_\tau(x^*, x^{\tau-1})$ is calculated as:

$$A_\tau(x^*, x^{\tau-1}) = \min\left(1, \frac{p(x^*)q(x^{\tau-1}|x^*)}{p(x^{\tau-1})q(x^*|x^{\tau-1})}\right) \tag{2.18}$$

From this acceptance rate, a critical point can be observed that the constructed Markov Chain is balanced, which is:

$$\begin{aligned} p(x)q(x^*|x)A_\tau(x^*, x) &= \min\left(p(x)q(x^*|x), p(x^*)q(x|x^*)\right) \\ &= \min\left(p(x^*)q(x|x^*), p(x)q(x^*|x)\right) \\ &= p(x^*)q(x|x^*)A_\tau(x, x^*) \end{aligned} \tag{2.19}$$

There is no doubt the choice of proposal distribution $q(\cdot)$ would heavily influence the acceptance ratio. Low valued acceptance ratio would result in inefficient inference. This issue leads to the following Gibbs sampling part.

2.2.2 Gibbs Sampling

The Gibbs sampling is a special case of the M-H algorithm, which is proposed to cater for the proposal distribution's choice. Consider n random variables $\boldsymbol{x} = (x_1, \cdots, x_n)$, each step of the Gibbs sampling would sample the variables $\{x_i, i \in \{1, \cdots, n\}\}$ from its conditional posterior distribution given the other variables $(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$ in a sequential or random order. In step

τ, the sampling proceeds as follows:

$$\begin{aligned} x_1^\tau \sim &\quad p(x_1|x_2^{\tau-1}, \cdots, x_n^{\tau-1}) \\ x_2^\tau \sim &\quad p(x_2|x_1^\tau, x_3^{\tau-1}, \cdots, x_n^{\tau-1}) \\ &\quad \vdots \\ x_n^\tau \sim &\quad p(x_n|x_1^\tau, \cdots, x_{n-1}^\tau) \end{aligned} \tag{2.20}$$

where $(x_1^\tau, \cdots, x_n^\tau)$ denotes the variables in step τ.

It seems the conditional probability in Eq. (2.20) needs the full joint distribution of the whole variables $\boldsymbol{x}$; however, due to the factorization theorem, only s small subset of the variables (the variables' neighbours) are usually needed for each of $\{x_i, i \in \{1, \cdots, n\}\}$. Also, while taking this conditional distribution as the proposal distribution for the M-H algorithm, the acceptance ratio is obtained as $(\forall i \in \{1, \cdots, n\})$:

$$\begin{aligned} A_\tau(x^*, x^{\tau-1}) =& \min\left(1, \frac{p(x^*)q(x^{\tau-1}|x^*)}{p(x^{\tau-1})q(x^*|x^{\tau-1})}\right) \\ =& \min\left(1, \frac{p(x_i^*|\boldsymbol{x}_{\setminus i})p(x_i^{\tau-1}|\boldsymbol{x}_{\setminus i})}{p(x_i^{\tau-1}|\boldsymbol{x}_{\setminus i})p(x_i^*|\boldsymbol{x}_{\setminus i})}\right) \\ =& 1. \end{aligned} \tag{2.21}$$

Eq. (2.21) shows that every sample of the Gibbs sampling is accepted. Thus, the steps in Eq. (2.20) can be regarded as the M-H algorithm with acceptance ratio equal to 1.

The M-H algorithm and Gibbs sampling are the standard sampling method in inferencing the model. Apart from these two, there are numerous sampling methods in improving its efficiency, such as auxiliary sampling (that is, adding extra variables to facilitate efficient sampling), collapse sampling (that is, marginalizing out some hidden variables), block sampling (that is, sampling a bunch of variables together), and slice sampling (that is, adding a uniform variable to restrain the component number in the nonparametric Bayesian prior). Interested readers in these methods are encouraged to refer to (Andrieu, De Freitas, Doucet & Jordan 2003)(Bishop et al. 2006) for further understanding.

2.3 Preliminaries & Literature Review of Relational Models

The probabilistic graphical model has provided powerful tools in analyzing the relational data. Introductions are specially given to the latent class model and latent feature model for the basis of the coupling relations learning.

2.3.1 Latent Class model

The simplest case of the latent class model can be referred to as the Gaussian Mixture Model, in which each data point is assumed to be generated from one of the Gaussian distributions in these mixtures. The corresponding task is to infer each data point's indicator to the Gaussian distribution and the parameters of these Gaussian distributions. While extended to the relational data, the infinite relational model (Kemp, Tenenbaum, Griffiths, Yamada & Ueda 2006) is one benchmark model. Since its assumption of one node possesses only one role is usually insufficient for modelling the real world, its advanced derivative of the mixed-membership stochastic blockmodel is introduced here.

The mixed-membership stochastic blockmodel (MMSB) (Airoldi, Blei, Fienberg & Xing 2008) aims to model each node's individual mixed-membership distribution. In MMSB, each interaction e_{ij} corresponds to two membership indicators: s_{ij} from the sender i and r_{ij} to the receiver j. (without loss of generality, it is assumed $s_{ij} = k$, $r_{ij} = l$). The interaction's value is determined by the compatibility of two corresponding communities k and l. Figure 2.1 shows the graphical model, and the detailed generative process can be described as:

- $\forall\{k, l\} \in \mathcal{N} > 0$, draw the communities' compatibility values $W_{k,l} \sim \text{Beta}(\lambda_1, \lambda_2)$, k, l refer to the communities' index
- $\forall i \in \{1, \cdots, n\}$, draw node i's mixed-membership distribution $\pi_i \sim \text{Dirichlet}(\beta)$

- $\forall\{i,j\} \in \{1,\cdots,n\}^2$, for interaction e_{ij}
  - sender's membership indicator $s_{ij} \sim \text{Multinomial}(\pi_i)$
  - receiver's membership indicator $r_{ij} \sim \text{Multinomial}(\pi_j)$
  - the interaction $e_{ij} \sim \text{Bernoulli}(W_{s_{ij},r_{ij}})$

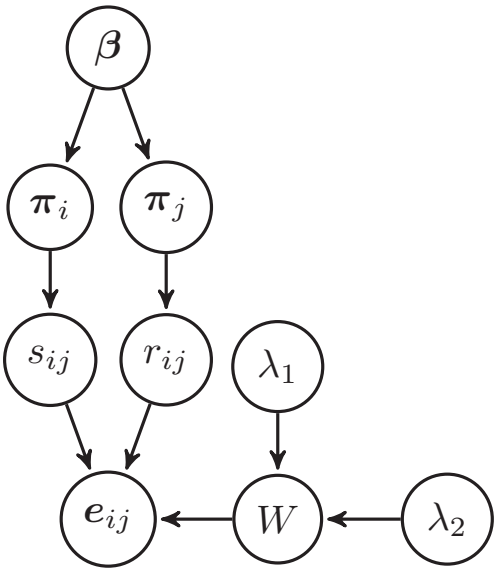

Figure 2.1: The mixed-membership stochastic blockmodel (MMSB) Model

It should be noted that each $\boldsymbol{\pi}_i$ is responsible for generating both the sender's labels $\{s_{ij}\}_{j=1}^n$ from node i and the receiver's labels $\{r_{ji}\}_{j=1}^n$ to node i.

W is the communities' compatibility matrix as described previously. The prior $P(W)$ is element-wise Beta distributed, which is conjugate to the Bernoulli distribution $P(e_{ij}|W, s_{ij}, r_{ij})$. Therefore, a marginal distribution of $P(e_{ij})$, i.e., $\int_W p(e_{ij}|W)p(W)d(W)$ can be obtained analytically, and hence there is no need to explicitly sample the values of W.

2.3.2 Latent Feature Model

The latent feature model (LFM) (Miller, Jordan & Griffiths 2009) provides an alternative point of view as to how to model the node's binary latent features, instead of its latent classes. Compared to the LCM, the LFM uses all of one node's occupied latent features in generating all its relations with other nodes. In other words, the LFM model assumes each node has links with others under one single binary vector, and this vector indicates the hidden features to which it occupies.

As shown in the graphical model of Figure 2.2, the detailed generative process can be described as:

- $\forall\{k,l\} \in \mathcal{N} > 0$, draw the communities' compatibility values $W_{k,l} \sim$ Normal$(0,1)$
- $\forall i \in \{1,\cdots,n\}$, draw node i's stick-breaking representation $\pi_i \sim$ Dirichlet(β)
- $\forall i \in \{1,\cdots,n\}$, draw node i's binary latent feature vector $z_i \sim$ Bernoulli(π_i)
- $\forall\{i,j\} \in \{1,\cdots,n\}^2$, for link data e_{ij}
  - the link data $e_{ij} \sim$ Bernoulli$(\frac{1}{1+\exp(-z_i W z_j)})$

2.3.3 Literature Review of the relational models

A detail literature review on the current state-of-the-art in relational models is provided here, including the latent class model (LCM), the latent feature model (LFM) and some other variants such as latent hierarchical model, dynamic relational models. For sufficient elaborations of these models, the interested readers are referred to (Schmidt & Morup 2013).

In general, the stochastic blockmodel (Nowicki & Snijders 2001) represents the baseline model in the relational model literature, in which it assumes that each node has a latent variable that directly represents its community

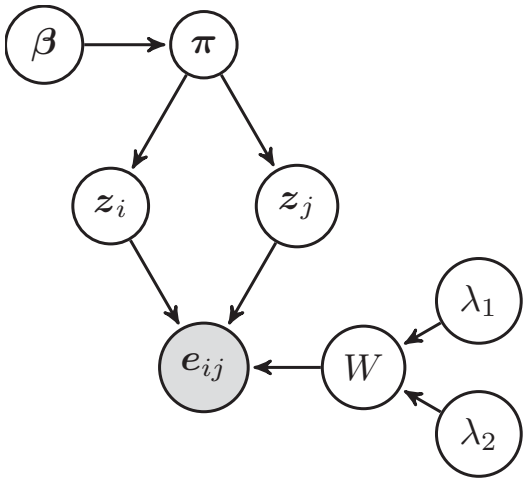

Figure 2.2: The graphical model for latent feature model (LFM).

membership. Each of the fixed number of communities is associated with a weight, and the whole weight vector can be seen as a draw from a K-dimensional Dirichlet distribution. Naturally, the community memberships are realized from the multinomial distribution parameterized by this weight vector. The binary link data between two nodes is determined by the communities to which they belong. This model has been extended to an infinite K community, i.e., infinite relational model (IRM) (Kemp et al. 2006) where the Dirichlet distribution has been replaced by a Dirichlet process.

Latent Class Model

On the LCM front, the classical approach is the mixed-membership stochastic blockmodel (MMSB) which enables each node to associate with multiple membership indicators, and an interaction is formed using one of these indicators, details of which are elaborated on in Section 2.3.1. Several representative works can be categorized into this LCM framework, including the infinite relational model. An attempt is made here to analyze them according to their various concentrations.

On the link data generation, the classic models include (Hofman & Wiggins 2008), which is to simplify the communities' compatibilities into 2 cases:

within-group probability (i.e. the membership indicators are the same) and between-group probability (i.e. the membership indicators are different). (Mørup & Schmidt 2012) further uses different within-group probabilities to generate the link, while keeping the between-group probability the same. Actually, the above two can be regarded as simplifications on treating the diagonal and off-diagonal values of the compatibility matrix. (Ishiguro, Ueda & Sawada 2012) uses an extra hidden noisy-indicator variable to handle the noisy relations.

On various extensions of MMSB, (Koutsourelakis & Eliassi-Rad 2008) extends the MMSB into the infinite communities case, by using the Chinese Restaurant Process to indicate the partition. (Ho, Parikh & Xing 2012) uses the nested Chinese Restaurant Process (Blei, Griffiths & Jordan 2010) to build the communities' hierarchical structure. (Kim et al. 2012) incorporates the node's attribute information into its membership indicator construction in MMSB.

Latent Feature Model

A representative work for the LFM is the latent feature relational model (LFRM) (Miller et al. 2009), which utilizes a latent feature matrix and a corresponding link generative function to define the model, as elaborated in detail in Section 2.3.2. To account for the variable number of features associated with each node, it uses the Indian Buffet Process (Griffiths & Ghahramani 2011)(Teh et al. 2007) as a prior. In the LFM, the compatibility matrix values may be negative for the generation of link data, that is, it prohibits the relation.

Several extensions on the LFM includes the max-margin latent feature relational model (Med-LFRM) (Zhu 2012), who uses the maximum entropy discrimination (MED) (Jebara 2004) technique to minimize the hinge loss, which is to measure the quality of link prediction. The infinite latent attribute (ILA) model (Palla, Ghahramani & Knowles 2012) uses a Dirichlet Process to construct a sub-structure within each feature and all the features

are utilized through the LFRM model. The infinite multiple-membership relational model (Morup, Schmidt & Hansen 2011) tries to alleviate the computational cost by assuming the link is generated independently given the (multiple) features that both of the two nodes occupy. (Knowles & Ghahramani 2007)(Hoff 2009) separately model the present indicator and the contribution of features, which leads to a nonparametric sparsity promoting prior.

Dynamic case

Like any data modeling problem, interaction data can also change across time, therefore, the dynamic extensions are found in both the LCM and LFM frameworks. Examples such as work on the time-varied relational model, for instance, the stochastic blockmodel is used to capture the evolving community's behavior across time (Yang, Chi, Zhu, Gong & Jin 2011), which is addressed in (Ishiguro, Iwata, Ueda & Tenenbaum 2010) by incorporating a time-varied Infinite Relational model. (Sarkar & Moore 2005)(Sarkar, Siddiqi & Gordon 2007) describe the time dependency by using Gaussian linear motion models. The dynamic relational infinite feature model (DRIFT) (Foulds, DuBois, Asuncion, Butts & Smyth 2011), which employs an independent Markov dynamic transition matrix to correlate consecutive time interaction data, is a natural extension of the latent feature relational model (LFRM). Latent feature propagation (LFP) (Heaukulani & Ghahramani 2013) directly integrates observed interactions, rather than the latent feature matrix, in the current time to model the distribution of latent features at the next time stamp. On the dynamic setting of mixed-membership stochastic blockmodel (MMSB), (Fu, Song & Xing 2009)(Xing, Fu & Song 2010) place a "parameter" (the mean) dependent Gaussian distribution to consider the time correlation, while (Ho, Song & Xing 2011) considers hierarchical communities modeling that evolves. However, as both of these two models require pre-definition of the number of communities, additional techniques, such as Cross-Validation, are necessary when choosing the number of communities.

Furthermore, their implicit description of the time dependency may not be sufficiently intuitive.

Other extensions

Apart from the above extensions, many of the networks are believed to be under a hierarchical setting. In modelling the hierarchical network structure, (Clauset, Moore & Newman 2008) defines a uniform prior over all binary trees, in which the probability of generating a link between two nodes is parameterized at the level of their nearest common ancestor in the binary tree. (Herlau, Morup, Schmidt & Hansen 2012) replaces the uniform prior over binary tree with a uniform prior over multifurcating trees and the leaves of the trees is generated from a Chinese Restaurant Process prior. (Roy, Kemp, Mansinghka & Tenenbaum 2007) assumes that each edge in the tree has an associated weight that defined the propensity in which the network compiles with the given split. The Mondrian Process (Roy & Teh 2009) splits the set of nodes into two parts at the first step, and continues this splitting until a stopping criterion is met. This can be regarded as a distribution over a k-dimensional tree. (Schmidt, Herlau & Mørup 2013) uses the Gibbs fragmentation tree as a prior over multifurcating trees, which is closely related to the two parameter nested Chinese Restaurant Process (Aldous 1985).

On utilizing multiple data source, (Kemp et al. 2006) and (Xu, Tresp, Yu & Kriegel 2006) extend the IRM model to both model dyadic relationships as well as side information, such that the partition of the nodes and the available side-information are equal. (Miller et al. 2009) incorporates the node attribute information directly for modelling the link, while (Kim et al. 2012) tries to use this information to uncover the hidden structure of the network.

2.4 Literature Review of Coupling Relations Learning

The real applications of data analysis often exhibit strong coupling relations and heterogeneities between the objects. This can not be intuitively simplified by the independently and identically distributed (i.i.d.) assumption. Correspondingly, studying this coupling relations (or non-i.i.d. learning) has emerged as a crucial problem, which is partially but not systematically proposed by (Zhang, Song, Gretton & Smola 2009)(Guo & Shi 2011)(Mohri & Rostamizadeh 2009)(Mohri & Rostamizadeh 2010) (Chazottes, Collet, Külske & Redig 2007).

The coupling relations learning, as well as non-i.i.d. learning, are particularly driven by the issues of (Cao 2013)(Wang & Cao 2012)(Cao, Dai & Zhou 2009)(Cao, Zhang & Zhou 2008) (Cao, Luo & Zhang 2009)(Cao 2010). These works discuss the complexities of the problem from several aspects, including *openness* (exchange between energy, information and materials), *scalability* (big data issue), *dynamic* (time related), *sufficient* (fruitful profile information), *heterogeneity* (different representation forms of the data). Under this emergent need, the coupling relations are systematically proposed by (Cao 2014).

There have been several case studies in exploring these coupling relationships. (Wang, Cao, Wang, Li, Wei & Ou 2011) and (Wang, She & Cao 2013*a*) are focusing on the coupled similarity description, which in turn has improved the performance comparing to the previous methods. (Wang, She & Cao 2013*b*) further uses the idea into the ensemble learning, based on existing clustering or classification results. Other interesting extensions include (Li, Wang, Cao & Yu n.d.), who uses a coupled pattern mining to deal with the imbalanced data problem, (Yu, Wang, Gao, Cao & Chen 2013), who considers the coupling relations in recommendation system and (Cheng, Miao, Wang & Cao 2013) by applying this coupling idea in the document clustering problem. However, none of these works are focusing on the prob-

abilistic relational model perspective, and the analysis of social network is also absent.

2.4.1 Limitation of the coupling Relational Learning

As should be emphasized, the current work in the coupling relational learning does not consider the communities' behaviour. Especially, the influence of communities and the simultaneous behaviour within the same communities is largely been overlooked. In terms of this disadvantages, this thesis has focused on using the community relational modelling method to simultaneously learn the coupling relations. More specifically, the major work of the thesis can be divided into three parts.

Nodes' dynamic coupling behaviour with different communities. As the serial coupling relation focus on the dynamic behaviour, Chapter 3 discusses this serial coupling relation from the communities' point of view. The nodes' mixed-membership distribution is studied as the basis for the this study.

The within communities' coupling behaviour As the casual and exclusive coupling relations characterize the positive and negative effects on each other, their potential implications within the communities are of high importance. Chapter 4 uses the concept of Copula function to describe these casual and exclusive coupling relations, which receives quite interesting results.

The coupling interaction between the nodes' information and the communities As the dependent coupling relation is to describe the potential connection between two objects, Chapter 5 efficiently discusses on the coupling relation between the nodes' information and the communities.

Chapter 3

Dynamic Infinite Mixed-Membership Stochastic Blockmodel

3.1 Introduction

Networking applications with dynamic settings (i.e. networks observed over time) are widely seen in real world environments, such as link prediction and community detection in social networks, social media interactions, capital market movements, and recommender systems. In this dynamic case, the objects usually have strong correlations between the consecutive times, which we denote them as the coupling relations along the time (or **serial coupling relations**.) A deep understanding of such dynamic network mechanisms relies on latent relation analysis and latent variable modeling of dynamic network interactions and structures. This presents both challenges and opportunities to existing learning theories. The intricacy associated with the time-varying attributes makes learning and inference a difficult task, while at the same time, one can explore the evolutionary behavior of a network structure more realistically in this time varying setting. The various dynamic characteristics of such a network can therefore be revealed in real

application.

A number of researchers have recently attempted to address this issue. Some notable earlier examples include stochastic blockmodel (Nowicki & Snijders 2001) and its infinite community case infinite relational model (IRM) (Kemp et al. 2006) where the aim is to partition a network of nodes into different groups based on their pairwise and directional binary interactions. It was extended by (Yang et al. 2011) to infer the evolving community's behavior over time. Their work assumes that a fixed number of K communities exists to which one node can potentially belong. However, in many applications, an accurate guess of K beforehand may be impractical and its value may also vary during the time stamps.

The dynamic infinite relational model (dIRM) (Ishiguro et al. 2010) is an alternative way to address the same problem, where K can be inferred from the data itself. However, just as in (Kemp et al. 2006), its drawback is that the model assumes each node i must belong to only one single community. Therefore, an interaction between nodes i and j can only be determined from their community indicators. This approach can be inflexible in many scenarios, such as the monastery example depicted in (Airoldi et al. 2008), where one monk can belong to different communities. To this end, the authors in (Airoldi et al. 2008) introduced the concept of mixed-membership, where they assume each node i might belong to multiple communities. The membership indicators of one's interaction are no longer a fixed value of special community. Instead, they are sampled from the nodes' mixed-membership distributions.

The above-mentioned works address some aspects (infinite, dynamic, mixed-membership and data-driven inference) of relational modeling respectively. An emergent need is to effectively unify these models to provide a flexible and generalised framework which can encapsulate the advantages of most of these works and address multiple aspects of complexities in one model. This is certainly not an easy thing to do because of the need to understand the relations between aspects and build a seamless approach to aggregate the

challenges. Accordingly, we propose the dynamic infinite mixed-membership stochastic blockmodel (DIM3).

DIM3 has the following features: firstly, it allows a network to have an infinite number of latent communities; secondly, it allows mixed-membership associated with each node; thirdly, the model extends to dynamic settings and the number of communities varies with the time; lastly, it is apparent that in many social networking applications, a node's membership may become consistent (i.e. unchanged) over consecutive time stamps, for example, a person's opinion of a peer is more likely to be consistent in two consecutive time stamps.

To model this persistence, we devise two different implementations. The first is to have a single mixed-membership distribution for each node at different time intervals. The persistence factor is dependent on the statistics of each node's interactions with the rest of the nodes. The second implementation is to allow a set of mixed-membership distributions to associate with each node, and they are time-invariant. The number of elements in the set varies non-parametrically, similar to (Fox et al. 2008). The persistence factor is dependent on the value of the membership indicator at the previous time stamp.

Two effective sampling algorithms are consequently designed for our proposed models, using either the Gibbs or Slice sampling technique for efficient model inference. Their convergence behavior and mixing rate are analyzed and displayed in the first part of the experiment. In the experimental analysis, we show that we can assess the nodes' position in the network and their developing trends, predict unknown links according to the current structure, understand the network structure and identify the change point. The techniques proposed can be used for forecasting the political tendencies of senators (Ho et al. 2011), predicting the function of a protein in biology (Xing et al. 2010), and tracking authors' community cooperation in academic circles (Heaukulani & Ghahramani 2013), etc.

The rest of the chapter is organised as follows. Section 3.2 details our

main framework and explains how it can incorporate infinite communities in a dynamic setting. The inference schemes for the two models are detailed in Section 3.3. In Section 3.4, we show the experimental results of the proposed models using both synthetic and real-world social network data. Conclusions and future works can be found in Section 3.5.

3.2 The dynamic infinite mixed-membership stochastic blockmodel (DIM3)

3.2.1 The general settings

In our DIM3 model, we allow each node's membership indicators to change across time. Additionally, it is imperative that these indicators should contain the time-persistence property with past values, through which the reality of social behavior can be reflected. Here, we use the strategy of incorporating a sticky parameter κ into the mixed-membership distributions to approach this issue (Fox et al. 2008)(Fox, Sudderth, Jordan & Willsky 2011*a*). Different detailed designs are proposed for the following two models, however, the idea that the current mixed-membership distributions are influenced by the corresponding distributions at the previous time is shared.

Once the current mixed-membership distributions have been selected, the interaction data is generated in the same way as MMSB. Thus, this paper is focused on the details of mixed-membership distribution constructions following the main route of the Hierarchical Dirichlet Process (HDP) (Teh et al. 2006). Also, we should note that the intermediate variable $\boldsymbol{\beta}$ is identical for both models, representing the "significance" of all the communities across time, and its construction is the same as the stick-breaking construction in Section 2.3.1.

3.2.2 The mixture time variant (MTV) Model

In Figure 3.1, we illustrate the graphical model of the MTV model. Here we only show all the variables involved for time t, and omit the other times, where the structure is identical at any other time $\tau \neq t$.

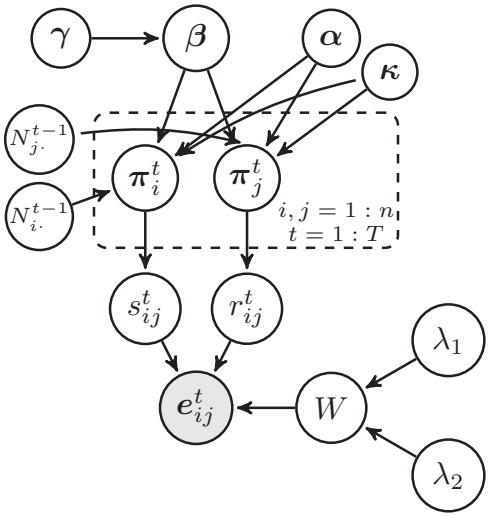

Figure 3.1: The mixture time variant (MTV) Model

Let us focus on the mixed-membership distribution's construction in the MTV model, which is:

$$\boldsymbol{\pi}_i^t \sim DP\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \frac{\kappa}{2n} \cdot \sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k}{\alpha + \kappa}\right) \tag{3.1}$$

$$s_{ij}^t \sim \boldsymbol{\pi}_i^t, r_{ij}^t \sim \boldsymbol{\pi}_j^t, \forall i, j \in \mathcal{N}, t \geq 1. \tag{3.2}$$

The mixed-membership distribution $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$ is sampled from the Dirichlet Process with a concentration parameter $(\alpha + \kappa)$ and a base measure $\frac{\alpha\boldsymbol{\beta}+\frac{\kappa}{2n}\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k}{\alpha+\kappa}$. There will be $N \times T$ of these distributions. They jointly describe each node's activities.

In the base measure, the introduced sticky parameter κ stands for each node's time influence on its mixed-membership distribution. In other words, we assume that each node's mixed-membership distribution at time t will be

largely influenced by its activities at time $t-1$. This is reflected in the hidden label's multinomial distribution whereby the previous explicit activities will occupy a fixed proportion $\frac{\kappa}{\alpha+\kappa}$ of the current distribution. The larger the value of κ, the more weight the activities at $t-1$ will have at time t.

As our method is largely based on the HDP framework, we will use the popular "Chinese Restaurant Franchise (CRF)" (Teh et al. 2006)(Fox et al. 2008) analogy to explain our model. Using the CRF analogy, the mixed-membership distribution associated with a node i at time t can be seen as a restaurant $\boldsymbol{\pi}_i^t$, with its dishes representing the communities. If a customer s_{ij}^t(or r_{ji}^t) eats the dish k at the i^{th} restaurant at time t, then $s_{ij}^t(r_{ji}^t) = k$. For all $t > 1$, the restaurant $\boldsymbol{\pi}_i^t$ will have its own specials on the dishes served, representing the "sticky" configuration in the graphical model. In contrast to the sticky HDP-HMM (Fox et al. 2008) approach, which places emphasis on one dish only, we allow multiple specials in our work, where the weight of each special dish is adjusted according to the number of dishes served at this restaurant at time $t-1$, i.e., $\frac{\kappa}{2n}\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k$. Therefore, we can ensure that the special dishes are served persistently across time in the same restaurant.

3.2.3 The mixture time invariant (MTI) Model

We show the MTI model in Figure 3.2. Here we only show the interaction e_{ij}^1 and omit the other interactions, whose structure is directly derived.

The $\boldsymbol{\beta}$ in the MTI model is identical to that in the MTV model, and we sample the mixed-membership distribution and membership indicators as follows:

$$\boldsymbol{\pi}_i^{(k)} \sim DP\left(\alpha+\kappa, \frac{\alpha\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k}{\alpha+\kappa}\right), \forall i, k \in \mathcal{N}; \tag{3.3}$$

$$s_{ij}^t \sim \boldsymbol{\pi}_i^{(s_{ij}^{t-1})}, r_{ij}^t \sim \boldsymbol{\pi}_j^{(r_{ij}^{t-1})}, \forall i, j \in \mathcal{N}, t \geq 1. \tag{3.4}$$

We assign uninformative priors on sampling the initial membership indicators $\{s_{ij}^0, r_{ij}^0\}_{i,j}$, i.e., $\{s_{ij}^0, r_{ij}^0\}_{i,j}$ are sampled from a multinomial distribution, with each category having an equalized success probability. The dimension of this

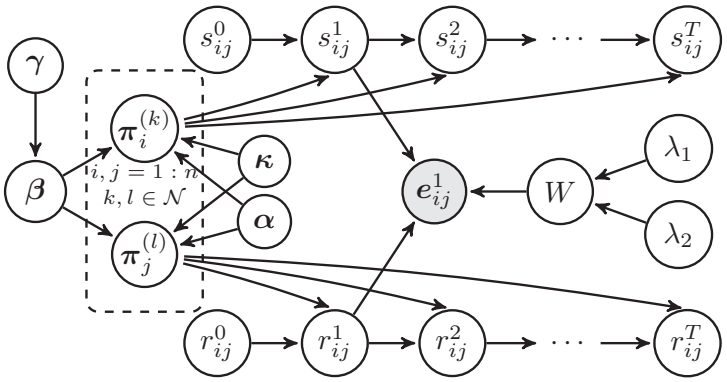

Figure 3.2: The mixture time invariant (MTI) Model

multinomial distribution is automatically adjusted according to the current number of communities in the model.

On each node's membership distribution, our MTI model is essentially a Sticky Hierarchical Dirichlet Process-Hidden Markov Model (HDP-HMM) (Fox et al. 2008)(Fox et al. 2011*b*)(Fox et al. 2011*a*). In this model, each node has a variable number of mixed-membership distributions associated with it, which may be infinite. At time $t \geq 2$, its membership indicator s_{ij}^t (or r_{ij}^t) is generated from $\boldsymbol{\pi}_i^{(s_{ij}^{t-1})}$ (or $\boldsymbol{\pi}_j^{(r_{ij}^{t-1})}$). To encourage persistence, each π_{ik} is generated from the corresponding $\boldsymbol{\beta}$, where κ is added to $\boldsymbol{\beta}$'s k^{th} component (Fox et al. 2008)(Fox et al. 2011*b*)(Fox et al. 2011*a*).

Returning to the CRF (Teh et al. 2006) analogy, we have $N \times \infty$ matrix, where its $(i,k)^{\text{th}}$ element refers to $\pi_i^{(k)}$, which can be seen as the weights of eating each of the available dishes. A customer s_{ij}^t(or r_{ji}^t) can therefore only travel between restaurants located at the i^{th} row of the matrix. When $\pi_i^{(k)}$'s k^{th} component is more likely to be larger, it means that the dish k is a special dish for restaurant k. Therefore, a customer at restaurant k at time $t-1$ is more likely to eat the same dish (i.e., k^{th} dish), and hence to stay at restaurant k at time t.

3.2.4 Discussion and comparison

As a brief comparison, we discuss the differences between the two models in the design of the time-persistence property. The MTV model allows the mixed-membership distribution itself to change over time stamps. However, there is only a single (but different) distribution for each node at each individual time stamp. The membership indicator of a node at time t is dependent on the "statistics" of all membership indicators of the same node at $t-1$ and $t+1$. With a larger value of the sticky parameter κ, the current mixed-membership distribution tends to be more similar to that of the previous time stamp.

In contrast, the MTI model requires the mixed-membership distributions to stay invariant over times. However, there may be an infinite number of possible distributions associated with each node, but often, due to a HDP prior, only a few distributions will be discovered. In this case, the membership indicator at the current time is dependent and more likely to have the same value as it had in the previous time stamp.

3.3 Inference

Two sampling schemes are implemented to complete the inference on the MTV model: standard Gibbs sampling and Slice-Efficient sampling, which both target the same posterior distribution.

3.3.1 Gibbs Sampling for the MTV model

The Gibbs Sampling scheme is largely based on (Teh et al. 2006). The variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables eating dish k as used in (Teh et al. 2006)(Fox et al. 2008) without counting the tables that are generated from the sticky portion, i.e., κN_{ik}^{t-1}. Note that we do not sample $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$, as it gets integrated out.

Sampling $\boldsymbol{\beta}$

$\boldsymbol{\beta}$ is the prior for all $\{\boldsymbol{\pi}_i^t\}s$, which can be thought of as the ratios between the community components for all communities. Its posterior distribution is obtained through the auxiliary variable $\hat{\boldsymbol{m}}$:

$$(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K, \boldsymbol{\beta}_\mu) \sim Dir(\hat{\boldsymbol{m}}_{\cdot 1}, \cdots, \hat{\boldsymbol{m}}_{\cdot K}, \gamma) \tag{3.5}$$

where its detail can be found in (Teh et al. 2006).

Sampling $\{s_{ij}^t\}_{n\times n}^{1:T}$, $\{r_{ij}^t\}_{n\times n}^{1:T}$

Each observation e_{ij}^t is sampled from a fixed Bernoulli distribution, where the Bernoulli's parameter is contained within the role-compatibility matrix W indexed (row and column) by a pair of corresponding membership indicators $\{s_{ij}^t, r_{ij}^t\}$. W.l.o.g., $\forall k, l \in \{1, \cdots, K+1\}$, the joint posterior probability of $(s_{ij}^t = k, r_{ij}^t = l)$ is:

$$\begin{aligned}
&\Pr(s_{ij}^t = k, r_{ij}^t = l | Z \backslash \{s_{ij}^t, r_{ij}^t\}, e, \boldsymbol{\beta}, \alpha, \lambda_1, \lambda_2, \kappa) \\
\propto &\Pr(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \\
&\cdot \prod_{l=1}^{2n} \Pr(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}) \\
&\cdot \Pr(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0 \neq i}, \{s_{j i_0}\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}) \\
&\cdot \prod_{l=1}^{2n} \Pr(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}) \\
&\cdot \Pr(e_{ij}^t | E \backslash \{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2)
\end{aligned} \tag{3.6}$$

The first two terms of Eq. (3.6):

$$\begin{aligned}
&\Pr(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \\
&\cdot \prod_{l=1}^{2n} \Pr(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}) \\
\propto &\frac{\Gamma(\alpha\boldsymbol{\beta}_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa)}{\Gamma(\alpha\boldsymbol{\beta}_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t})} \cdot \frac{\Gamma(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t,-s_{ij}^t})}{\Gamma(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa)} \\
&\cdot \begin{cases} \alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t}, & k \in \{1, \ldots, K\}; \\ \alpha\boldsymbol{\beta}_\mu, & k = K+1. \end{cases}
\end{aligned} \tag{3.7}$$

Here we should note that $N_{ik}^0 = 0$, $N_{ik}^{T+1} = 0$, $\forall i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$.

The following two terms of Eq. (3.6) are:

$$\begin{aligned}
&\Pr(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0 \neq i}, \{s_{j i_0}^t\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}) \\
&\cdot \prod_{l=1}^{2n} \Pr(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}) \\
\propto &\frac{\Gamma(\alpha\boldsymbol{\beta}_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa)}{\Gamma(\alpha\boldsymbol{\beta}_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t})} \cdot \frac{\Gamma(\alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t,-r_{ij}^t})}{\Gamma(\alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa)} \\
&\cdot \begin{cases} \alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t-1} + N_{jl}^{t,-r_{ij}^t}, & l \in \{1, \ldots, K\}; \\ \alpha\boldsymbol{\beta}_\mu, & l = K+1. \end{cases}
\end{aligned} \tag{3.8}$$

The last term, i.e. the likelihood term is calculated as that of:

$$\begin{aligned}
&\Pr(e_{ij}^t | E \backslash \{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2) \\
= &\begin{cases} \frac{n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} + \lambda_1}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & \boldsymbol{e}_{ij}^t = 1; \\ \frac{n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^t} + \lambda_2}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & \boldsymbol{e}_{ij}^t = 0. \end{cases}
\end{aligned} \tag{3.9}$$

Here $n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} = n_{k,l}^t - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l) = \sum_{i'j'} \mathbf{1}(s_{i'j'}^t = k, r_{i'j'}^t = l) - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)$, $n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} = n_{k,l}^{1,t} - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)\boldsymbol{e}_{ij}^t = \sum_{i'j':s_{i'j'}^t=k, r_{i'j'}^t=l} \boldsymbol{e}_{i'j'}^t - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)\boldsymbol{e}_{ij}^t$ and $n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^t} = n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} - n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t}$.

The detail derivation of Equation (3.7)(3.8)(3.9) are given in the supplementary material. Assuming the current sample of $\{s_{ij}^t, r_{ij}^t\}$ has values ranging between $1 \dots K$, we let the undiscovered (new) community be indexed by $K+1$. Then, to sample a pair (s_{ij}^t, r_{ij}^t) in question, we need to calculate all $(K+1)^2$ combinations of values for the pair.

Sampling $\hat{\boldsymbol{m}}$

Using the restaurant-table-dish analogy, we denote $\boldsymbol{m}_{ik}^t$ as the number of tables eating dish $k, \forall i, k, t$. This is related to the variable $\hat{\boldsymbol{m}}$ used in sampling $\boldsymbol{\beta}$, but also includes the counts of the "un-sticky" portion, i.e., $\alpha\boldsymbol{\beta}_k$.

The sampling of $\boldsymbol{m}_{ik}^t$ incorporates a similar strategy as (Teh et al. 2006)(Fox et al. 2008), which is independently distributed from:

$$\Pr(\boldsymbol{m}_{ik}^t = m | \alpha, \boldsymbol{\beta}_k, N_{ik}^{t-1}, \kappa) \propto S(N_{ik}^t, m)(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t-1})^m \tag{3.10}$$

Here $S(\cdot, \cdot)$ is the Stirling number of the first kind.

For each node, the ratio of generating new tables is the result of two factors: (1) Dirichlet prior with parameter $\{\alpha, \boldsymbol{\beta}\}$ and (2) the sticky configuration from membership indicators at $t-1$, i.e., κN_{ik}^{t-1}.

To sample $\boldsymbol{\beta}$, we need only include tables generated from the "un-sticky" portion, i.e., $\hat{\boldsymbol{m}}$, where each $\hat{\boldsymbol{m}}_{ik}^t$ can be obtained from a single Binomial draw:

$$\hat{\boldsymbol{m}}_{ik}^t \sim Binomial(\boldsymbol{m}_{ik}^t, \frac{\alpha\boldsymbol{\beta}_k}{\frac{\kappa}{2n}N_{ik}^{t-1} + \alpha\boldsymbol{\beta}_k}). \tag{3.11}$$

$$\hat{\boldsymbol{m}}_k = \sum_{i,t} \hat{\boldsymbol{m}}_{ik}^t. \tag{3.12}$$

3.3.2 Adapted Slice-Efficient Sampling for the MTV model

We also incorporate slice-efficient sampling (Kalli, Griffin & Walker 2011)(Walker 2007) to our model. The original sampling scheme was designed to sample the Dirichlet Process Mixture model. To adapt it to our framework, which is based on a HDP prior and also has pair-wise membership indicators, we use the auxiliary variables $U = \{u^t_{ij,s}, u^t_{ij,r}\}$ for each of the latent membership pairs $\{s^t_{ij}, r^t_{ij}\}$. Having the Us, we are able to limit the number of components in which $\boldsymbol{\pi}_i$ needs to be considered, which is otherwise infinite.

Under the slice-efficient sampling framework, the variables of interest are now extended to: $\boldsymbol{\pi}^t_i, \{u^t_{ij,r}, u^t_{ij,s}\}, \{s^t_{ij}, r^t_{ij}\}, \boldsymbol{\beta}, \boldsymbol{m}$:

Sampling $\boldsymbol{\pi}^t$

For each node $i = 1, \cdots, N; t = 1, \cdots, T$: we generate $\boldsymbol{\pi}^{'t}_i$ using the stick-breaking process (Ishwaran & James 2001), where each k^{th} component is generated using:

$\boldsymbol{\pi}^{'t}_{ik} \sim$beta$(\boldsymbol{\pi}^{'t}_{ik}; a^t_{ik}, b^t_{ik})$, where

$$\begin{aligned} a^t_{ik} &= \alpha\boldsymbol{\beta}_k + N^t_{ik} + \kappa N^{t-1}_{ik} \\ b^t_{ik} &= \alpha(1 - \sum_{l=1}^{k} \boldsymbol{\beta}_l) + N^t_{i,k_0>k} + \kappa N^{t-1}_{i,k^0>k} \end{aligned} \tag{3.13}$$

Here $\boldsymbol{\pi}^t_k = \boldsymbol{\pi}^{'t}_k \prod_{i=1}^{k-1}(1 - \boldsymbol{\pi}^{'t}_i)$.

Sampling $u^t_{ij,s}, u^t_{ij,r}, s^t_{ij}, r^t_{ij}$

We use $u^t_{ij,s} \sim U(0, \boldsymbol{\pi}^t_{is^t_{ij}})$, $u^t_{ij,r} \sim U(0, \boldsymbol{\pi}^t_{jr^t_{ij}})$. The hidden label subsequently obtained is then independently sampled from the finite candidates:

$$\begin{aligned} & P(s^t_{ij} = k, r^t_{ij} = l | Z, e^t_{ij}, \boldsymbol{\beta}, \alpha, \kappa, N, \boldsymbol{\pi}, u^t_{ij,s}, u^t_{ij,r})) \\ \propto & \mathbf{1}(\boldsymbol{\pi}^t_{ik} > u^t_{ij,s}) \cdot \mathbf{1}(\boldsymbol{\pi}^t_{jl} > u^t_{ij,r}) \\ & \cdot \prod_{l=1}^{2n} \Pr(z^{t+1}_{il} | z^t_{i\cdot}/s^t_{ij}, s^t_{ij} = k, \boldsymbol{\beta}, \alpha, \kappa, N^{t+1}_i) \\ & \cdot \prod_{l=1}^{2n} \Pr(z^{t+1}_{jl} | z^t_{j\cdot}/r^t_{ij}, r^t_{ij} = l, \boldsymbol{\beta}, \alpha, \kappa, N^{t+1}_j) \\ & \cdot \Pr(e^t_{ij} | E \backslash \{e^t_{ij}\}, s^t_{ij} = k, r^t_{ij} = l, \boldsymbol{Z} \backslash \{s^t_{ij}, r^t_{ij}\}, \lambda_1, \lambda_2) \end{aligned} \tag{3.14}$$

We refer to Eq. (3.7)(3.8)(3.9) for the detailed calculation of each term in Eq. (3.14).

Sampling $\boldsymbol{\beta}$

An obvious choice for the proposal distribution of $\boldsymbol{\beta}$ used in M-H is its prior $p(\boldsymbol{\beta}|\gamma) = \text{stick-breaking}(\gamma)$. However, this proposal can be non-informative, which results in a low acceptance rate. We sample $\boldsymbol{\beta}^*$ conditioned on an auxiliary variable $\hat{\boldsymbol{m}}$: $(\boldsymbol{\beta}^*_1, \cdots, \boldsymbol{\beta}^*_K, \boldsymbol{\beta}^*_{K+1}) \sim Dir(\hat{\boldsymbol{m}}_1, \cdots, \hat{\boldsymbol{m}}_K, \gamma)$, in order to increase the M-H's acceptance rate, where $\hat{\boldsymbol{m}}$ are sampled in accordance with the method proposed in Section 3.3.1 (Eq. (3.10)(3.11)(3.12)). However, instead of sampling $\boldsymbol{\beta}$ directly from $\boldsymbol{m}$ as in Section 3.3.1, we only use it for our proposal distribution, as we have explicitly sampled $\{\pi_i\}_{i=1}^n$. The acceptance ratio is hence (τ indexes the iteration time):

$$A(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(\tau)}) = \min(1, a) \tag{3.15}$$

$$a = \frac{\prod_{t,i} \left[\prod_{d=1}^{K+1} \Gamma(\alpha \boldsymbol{\beta}^{(\tau)}_d) \cdot [\pi^t_{id}]^{\alpha \boldsymbol{\beta}^*_d} \right]}{\prod_{t,i} \left[\prod_{d=1}^{K+1} \Gamma(\alpha \boldsymbol{\beta}^*_d) \cdot [\pi^t_{id}]^{\alpha \boldsymbol{\beta}^{(\tau)}_d} \right]} \cdot \frac{\prod_{d=1}^{K} \left[\boldsymbol{\beta}^{(\tau)}_d \right]^{\hat{\boldsymbol{m}}_d - \gamma}}{\prod_{d=1}^{K} [\boldsymbol{\beta}^*_d]^{\hat{\boldsymbol{m}}_d - \gamma}} \tag{3.16}$$

3.3.3 Hyper-parameter Sampling

The hyper-parameters involved in the MTV model are γ, α, κ. However, it is impossible to compute their posterior individually. Therefore, we place three prior distributions on some "combination" of the variables. A vague gamma prior $\mathcal{G}(1,1)$ is placed on both $\gamma, (\alpha+\kappa)$. A beta prior is placed on the ratio $\frac{\kappa}{\alpha+\kappa}$.

To sample γ value, since $\log(\gamma)$'s posterior distribution is log-concave, we use the Adaptive Rejection Sampling (ARS) method (Rasmussen 1999).

To sample $(\alpha+\kappa)$, we use the Auxiliary Variable Sampling (Teh et al. 2006), and this needs the auxiliary variable $\boldsymbol{m}$ in Eq. (3.10), as proposed in (Teh et al. 2006).

To sample $\frac{\kappa}{\alpha+\kappa}$, we place a vague beta prior $\mathcal{B}(1,1)$ on it, with a likelihood of $\{\boldsymbol{m}_{ik}^t - \hat{\boldsymbol{m}}_{ik}^t, \forall i,k,t>1\}$ in Eq. (3.11), the posterior is in an analytical and samplable form, thanks to its conjugate property.

3.3.4 Gibbs Sampling for the MTI model

The variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables eating dish k as used in (Teh et al. 2006)(Fox et al. 2008) without counting the tables generated from the sticky portion, i.e., κN_{ik}^{t-1}. As the hyper-parameters in the MTI model are quite similar to those in (Fox et al. 2011a), we do not present the hyper-parameters here. Interested readers can refer to (Fox et al. 2008)(Fox et al. 2011b)(Fox et al. 2011a) for the detailed implementation.

Sampling $\boldsymbol{\beta}$

$\boldsymbol{\beta}$'s sampling is the same as Eq. (1).

Samplinig s_{ij}^t, r_{ij}^t

The posterior probability of s_{ij}^t, r_{ij}^t is:

$$\begin{aligned}
&\Pr(s_{ij}=k, r_{ij}=l|\alpha,\boldsymbol{\beta},\kappa,\{N_{\cdot\cdot}^{(i)}\},\{N_{\cdot\cdot}^{(j)}\},\boldsymbol{e},\lambda_1,\lambda_2,Z)\\
\propto&\Pr(s_{ij}^t=k|\alpha,\boldsymbol{\beta},\kappa,N_{s_{ij}^{t-1}\cdot}^{(i)},s_{ij}^{t-1})\Pr(r_{ij}^t=l|\alpha,\boldsymbol{\beta},\kappa,N_{r_{ij}^{t-1}\cdot}^{(j)},r_{ij}^{t-1})\\
&\cdot\Pr(\boldsymbol{e}_{ij}^t|\boldsymbol{e}/\{\boldsymbol{e}_{ij}^t\},s_{ij}^t=k,r_{ij}^t=l,Z/\{s_{ij}^t,r_{ij}^t\},\lambda_1,\lambda_2)
\end{aligned} \tag{3.17}$$

The first term of Eq. (3.17) is:

$$\begin{aligned}
&\Pr(s_{ij}^t=k|\alpha,\boldsymbol{\beta},\kappa,N_{s_{ij}^{t-1}\cdot}^{(i)},s_{ij}^{t-1})\\
\propto&(\alpha\boldsymbol{\beta}_k+N_{s_{ij}^{t-1}k}^{(i)}+\kappa\delta(s_{ij}^{t-1},k))\cdot\\
&\left(\frac{\alpha\boldsymbol{\beta}_{s_{ij}^{t+1}}+N_{ks_{ij}^{t+1}}^{(i)}+k\delta(k,s_{ij}^{t+1})+\delta(k,s_{ij}^{t-1})\delta(k,s_{ij}^{t+1})}{\alpha+N_{k\cdot}^{(i)}+\kappa+\delta(s_{ij}^{t-1},k)}\right)
\end{aligned} \tag{3.18}$$

The second term of Eq. (3.17) is:

$$\begin{aligned}
&\Pr(r_{ij}^t=l|\alpha,\boldsymbol{\beta},\kappa,N_{r_{ij}^{t-1}\cdot}^{(j)},r_{ij}^{t-1})\\
\propto&(\alpha\boldsymbol{\beta}_l+N_{r_{ij}^{t-1}l}^{(j)}+\kappa\delta(r_{ij}^{t-1},l))\cdot\\
&\left(\frac{\alpha\boldsymbol{\beta}_{r_{ij}^{t+1}}+N_{lr_{ij}^{t+1}}^{(i)}+l\delta(l,r_{ij}^{t+1})+\delta(l,r_{ij}^{t-1})\delta(l,r_{ij}^{t+1})}{\alpha+N_{l\cdot}^{(i)}+\kappa+\delta(r_{ij}^{t-1},l)}\right)
\end{aligned} \tag{3.19}$$

The likelihood of $\Pr(\boldsymbol{e}_{ij}^t|\boldsymbol{e}/\{\boldsymbol{e}_{ij}^t\},s_{ij}^t=k,r_{ij}^t=l,Z/\{s_{ij}^t,r_{ij}^t\},\lambda_1,\lambda_2)$ is the same as Eq. (3.9).

Sampling $\hat{\boldsymbol{m}}$

$\hat{\boldsymbol{m}}$ is similar to that in the MTV model, however, it differs in the incorporation of κ.

$$\Pr(\boldsymbol{m}_{qk}^{(i)}=m|\alpha,\boldsymbol{\beta}_k,\kappa,N_{qk}^{(i)})\propto S(N_{qk}^{(i)},m)(\alpha\boldsymbol{\beta}_k+\kappa) \tag{3.20}$$

$$\hat{\boldsymbol{m}}_{qk}^{(i)}\sim\text{Binomial}(\boldsymbol{m}_{qk}^{(i)},\frac{\alpha\boldsymbol{\beta}_k}{\kappa+\alpha\boldsymbol{\beta}_k}) \tag{3.21}$$

$$\hat{\boldsymbol{m}}_{\cdot k} = \sum_{q,i} \hat{\boldsymbol{m}}_{qk}^{(i)} \tag{3.22}$$

3.3.5 Inference discussions

Both the Gibbs Sampling and Slice-Efficient Sampling are two feasible ways to accomplish our task. They have different pros and cons.

As mentioned previously, Gibbs Sampling in our MTV model integrates out the mixed-membership distribution $\{\boldsymbol{\pi}_i^t\}$. It is the "marginal approach" (Papaspiliopoulos & Roberts 2008). The property of community exchangeability makes it simple to implement. However, theoretically, the obtained samples mix slowly as the sampling of each label is dependent on other labels.

Slice-Efficient Sampling is one "conditional approach" (Kalli et al. 2011) while the membership indicators are independently sampled from $\{\boldsymbol{\pi}_i^t\}$. In each iteration, given $\{\boldsymbol{\pi}_i^t\}$ and the role-compatibility matrix W, we can parallelize the process of sampling membership indicators, which may help to improve the computation, especially when the number of nodes, i.e., N becomes larger, and the number of communities, i.e., k becomes smaller.

$$\begin{vmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.95 & 0.05 \\ 0.05 & 0 & 0.95 \end{vmatrix} \qquad \begin{vmatrix} 0.95 & 0.2 & 0 \\ 0.05 & 0.95 & 0.05 \\ 0.2 & 0 & 0.95 \end{vmatrix}$$

$$\begin{vmatrix} 0.05 & 0.95 & 0 \\ 0.05 & 0.05 & 0.95 \\ 0.95 & 0 & 0.05 \end{vmatrix} \qquad \begin{vmatrix} 0.05 & 0.95 & 0 \\ 0.2 & 0.05 & 0.95 \\ 0.95 & 0 & 0.2 \end{vmatrix}$$

Figure 3.3: Four Cases of the Compatibility Matrix.

3.4 Experiments

The performance of our DIM3 model is validated by experiments on synthetic datasets and several real-world datasets. On the synthetic datasets, we implement the finite-communities cases of our models as baseline algorithms, namely as f-MTV model and f-MTI model. On the real world dataset comparison, we individually implement the three benchmark models: the mixed-membership stochastic blockmodel (MMSB), the infinite relational model (IRM) and the latent feature relational model (LFRM) to the best of our understanding. Also, we compare the dynamic relational infinite feature model (DRIFT) with our models on real world datasets, and the source code is provided by (Foulds et al. 2011).

3.4.1 Synthetic Dataset

For the synthetic data generation, the variables are generated following (Ho et al. 2011). We use $N = 20, T = 3$, and hence E is a $20 \times 20 \times 3$ asymmetric and binary matrix. The parameters are set up such that the 20 nodes are equally partitioned into 4 groups. The ground-truth of the mixed-membership distributions for each of the groups are: $[0.8, 0.2, 0; 0, 0.8, 0.2; 0.1,$
$0.05, 0.85; 0.4, 0.4, 0.2]$.

We consider 4 different cases to fully assess DIM3 against the ground-truth; all lie in the 3-role compatibility matrix.

The detailed value of the role-compatibility matrix on these four cases is shown in Figure 3.3. Top left (Case 1): large diagonal values and small non-diagonal values. Top right (Case 2): large diagonal values and mediate non-diagonal values. Bottom left (Case 3): large non-diagonal values and small diagonal values. Bottom right (Case 4): small diagonal values and mediate non-diagonal values.

MCMC Analysis

The convergence behavior is tested in terms of two quantities: the cluster number K, i.e., the number of different values Z can take, and the deviance D of the estimated density (Kalli et al. 2011)(Papaspiliopoulos & Roberts 2008), which is defined as:

$$D = -2 \sum_{i,j,t} \log \left(\sum_{k,l} \frac{N_{ik}^t \cdot N_{jl}^t}{4n^2 T} p(e_{ij}^t | Z, \lambda_1, \lambda_2) \right) \tag{3.23}$$

In our MCMC stationary analysis, we run 5 independent Markov chains and discard the first half of the Markov chains as a burn-in. With the random partition of 3 initial classes as the starting point, $20,000$ iterations are conducted in our samplings.

The simulated chains satisfy standard convergence criteria, as the test was implemented using the CODA package (Plummer, Best, Cowles & Vines 2006). In Gelman and Rubin's diagnostics (Gelman & Rubin 1992), the value of Proportional Scale Reduction Factor (PSRF) is 1.09 (with upper C.I. 1.27) for k, 1.03 (with upper C.I. 1.09) for D in the Gibbs sampling, and 1.02 (with upper C.I. 1.06) for k, 1.02 (with upper C.I. 1.02) for D in Slice sampling. Geweke's convergence diagnostics (Geweke 1992) are also employed, with the proportion of the first 10% and last 50% of the chain as comparison. The corresponding z-scores are all in the interval $[-2.09, 0.85]$ for 5 chains. In addition, the stationarity and half-width tests of Heidelberg and Welch Diagnostic (Heidelberger & Welch 1981) were both passed in all the cases, with p-value higher than 0.05. Based on all these statistics, the Markov chain's stationarity can be safely ensured in our case.

The efficiency of the algorithms can be measured by estimating the integrated autocorrelation time τ for K and D. τ is a good performance indicator as it measures the statistical error of Monte Carlo approximation on a target function f. The smaller τ, the more efficient the algorithm.

(Kalli et al. 2011) used an estimator $\widehat{\tau}$ as:

$$\widehat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \widehat{\rho}_l \tag{3.24}$$

Here $\widehat{\rho}_l$ is the estimated autocorrelation at lag l and C is a cut-off point, which is defined as $C := \min\{l : |\widehat{\rho}_l| < 2/\sqrt{M}\}$, and M is the number of iterations.

Table 3.1: Integrated Autocorrelation Times Estimator $\widehat{\tau}$ for K

		K				
Sampling	γ \ α	0.1	0.3	0.5	1	2
MTV-g	0.1	177.2	93.65	26.91	50.21	11.24
	0.3	260.5	54.00	9.18	5.31	6.56
	0.5	1.83	8.33	7.54	3.95	5.24
	1.0	5.57	6.45	3.44	3.64	4.56
	2.0	4.30	2.87	3.35	2.98	3.28
MTV-s	0.1	248.6	90.63	161.3	9.58	17.69
	0.3	120.6	66.23	44.35	11.40	7.28
	0.5	18.99	27.27	6.08	8.76	10.40
	1.0	5.79	9.19	11.85	8.46	7.25
	2.0	3.17	8.41	5.35	5.48	5.05

We test the sampling efficiency of the MTV-g and the MTV-s on Case 1 with the same setting as (Papaspiliopoulos & Roberts 2008). Of the whole $20,000$ iterations, the first half of the samples is discarded as a burn-in and the remainder are thinned 1/20. We manually try different values of the hyper-parameters γ and α and show the integrated autocorrelation time estimator in Table 3.1 and Table 3.2. Although some outliers exist, we can see that there is a general trend that, with fixed α value, the autocorrelation function will decrease when the γ value increases. This same phenomenon happens on α while γ is fixed. This result confirms our empirical knowledge. The larger value of γ, α will help to discover more clusters, followed by a smaller autocorrelation function.

On the other hand, we confirm that the MTV-g and the MTV-s do not show much difference in the mixing rate of the Markov Chain as shown

Table 3.2: Integrated Autocorrelation Times Estimator $\widehat{\tau}$ for D

		D				
Sampling	γ \ α	0.1	0.3	0.5	1	2
MTV-g	0.1	358.8	148.3	23.94	84.75	4.31
	0.3	389.5	315.0	3.11	26.32	4.78
	0.5	2.88	79.34	90.93	3.17	3.82
	1.0	3.19	2.78	1.76	8.14	5.74
	2.0	95.48	1.91	3.29	8.74	6.55
MTV-s	0.1	8.67	59.90	57.57	1.87	3.70
	0.3	29.05	20.64	30.01	45.57	3.40
	0.5	39.66	3.87	5.30	3.17	5.83
	1.0	40.51	4.85	3.12	6.88	10.51
	2.0	25.54	34.82	4.61	35.61	12.68

in Table 3.1 and Table 3.2. As mentioned in the previous section, Slice sampling provides a mixed-membership distribution independent sampling scheme, which enjoys the time efficiency of parallel computing in one iteration. For large scale datasets, it is a feasible solution. In Gibbs sampling, parallel computing is impossible as the sampling variables are in a dependent sequence.

Figure 3.4 is the trace plot of the training log-likelihood against the iterations on Case 1. As we can see, the sampler in the MTI model converges to the high training log-likelihood region faster than the MTV model. Also, the MTI model reaches a higher training log-likelihood than the MTV model.

Further Performance

We will compare the models in terms of the Log-likelihood (in Figure 3.5); the average l_2 distance between the mixed-membership distributions and its ground-truth; and the l_2 distance between the posterior role-compatibility matrix and its ground-truth (in Table 3.3).

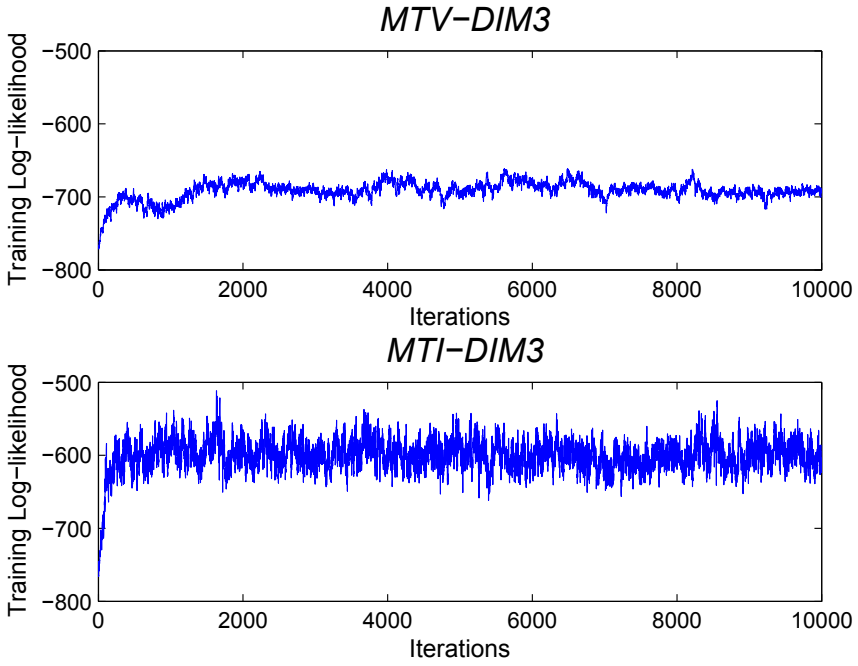

Figure 3.4: Top: the training log-likelihood trace plot on the MTV-g model. Bottom: the training log-likelihood trace plot on the MTI-g model.

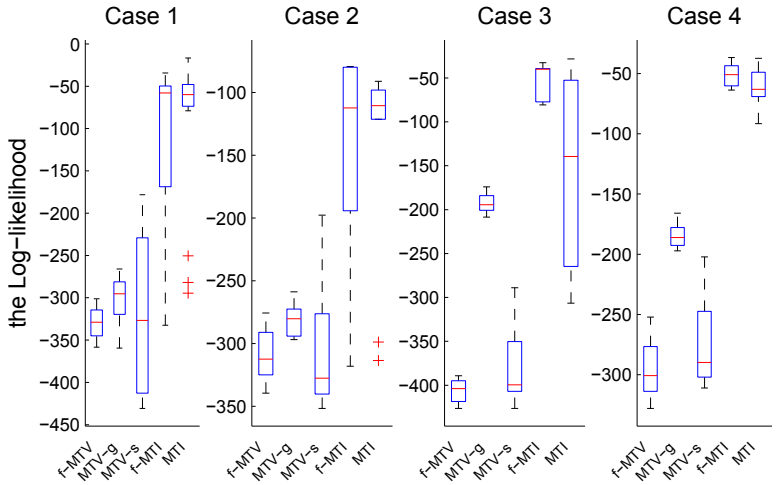

Figure 3.5: Log-likelihood Performance on all the four cases

Table 3.3: Average l_2 Distance to the Ground-truth

	Case	f-MTV	MTV-g	MTV-s	f-MTI	MTI	MMSB
Role-Compatibility	1	0.239	0.243	0.259	0.114	**0.086**	0.271
	2	0.206	0.225	0.240	**0.195**	0.204	0.285
	3	0.134	0.201	0.246	0.117	**0.087**	0.280
	4	**0.195**	0.214	0.267	0.220	0.219	0.246
	large size	0.220	0.239	0.215	0.142	**0.059**	0.237
Mixed-membership	1	0.366	0.384	0.403	0.199	**0.191**	0.411
	2	0.355	0.355	0.319	**0.207**	0.227	0.398
	3	0.278	0.289	0.589	0.208	**0.187**	0.329
	4	0.258	0.285	0.277	0.192	**0.182**	0.310
	large size	0.243	0.316	0.246	0.147	**0.120**	0.296

From the log-likelihood comparison in Figure 3.5, we can see that the MTI model performs better than the MTV model in general. On the average l_2 distance to the ground-truth performance, the MTI model also performs better. The superiority of the MTI model's performance over that of the MTV model is within our expectation, as the MTI model describes the membership indicator's time consistency more accurately (i.e. integrating the sticky parameter κ on the specific membership indicator, rather than the mixed-membership distribution). Also, the hidden Markov property enables the MTI model to categorize membership indicators into the same mixed-membership distributions based on its previous value. This seems to be a more effective way than the time-based grouping in the MTV model. However, in situations where there are dramatic changes amongst the membership distributions over times, then the MTI model will not respond well, i.e., the MTV model is much more effective and robust under these settings. In addition, the assumption that there exist different membership distributions at each time instance makes it possible to parallelise MTV model to some extent and making it suitable for dealing with large-scale problems.

Here we compare the computational complexity (Running Time) of the

Table 3.4: Running Time (Seconds per iteration)

N.	f-MTV	MTV-g	MTV-s	f-MTI	MTI	p-MTV-s[1]
20	0.20	0.28	0.23	0.15	0.31	0.29
50	1.03	1.52	1.29	0.95	1.91	1.79
100	3.69	5.76	4.81	3.74	7.49	5.06
200	15.61	24.17	19.87	15.82	30.19	21.64
500	106.96	154.45	119.82	105.61	202.09	132.43
1000	493.44	888.86	642.28	597.29	1102.90	393.24

[1] p-MTV-s denotes the parallel implementation of the MTV-s inference.

models in one iteration, with K discovered communities and show the results in Table 3.4. We discuss the MTV-g model and the MTV-s model as an instance. In the MTV-g model, the number of variables to be sampled is $(2K+2n^2T)$, while a total of $(2K+4n^2T+nT)$ variables are sampled in the MTV-s model. However, the posterior calculation of Z in the MTV-s model can be directly obtained from the mixed-membership distribution, while we need to calculate the ratio for each of Z in the MTV-g model. Also, the U value at each time can be sampled in one operation as its independency in the MTV-s model. The result is that the MTV-s model runs faster than the MTV-g model, which is in accordance with our assumption.

Also, we have tried a parallel implementation on the slice variables $\{u^t_{ij,s}, u^t_{ij,r}\}_{i,j,t}$'s in the MTV-s model. During each iteration, these slice variables are partitioned into 4 parts (as our machine has 4 cores) and sampled independently, while other variables are still sampled in a sequence. Its corresponding running time is exhibited in the last column of Table 3.4. As we can see, our parallel design cost even more time when the dataset size is small ($N \leq 500$). This may due to the time spent on transferring the variables. However, it needs less time when the dataset size is larger ($N > 500$). This verifies that our parallel slice sampling method is a promising way in achieving large scalability.

Figure 3.6: Training Log-likelihood Performance (95% Confidence Interval = Mean $\mp\times$Standard Deviation)

Dataset	MTV-g	MTV-s	MTI	
Kapferer	-673.7 ∓ 15.9	-698.9 ∓ 15.2	$\mathbf{-501.5 \mp 0.0}$	
Sampson	-347.6 ∓ 23.4	-350.4 ∓ 22.2	$\mathbf{-242.0 \mp 0.0}$	
Student-net	-1054.4 ∓ 48.5	-1059.3 ∓ 46.2	$\mathbf{-594.3 \mp 0.0}$	
Enron	-2274.2 ∓ 25.6	-2154.4 ∓ 43.3	$\mathbf{-1335.7 \mp 17.1}$	
Senator	-897.3 ∓ 16.2	-887.4 ∓ 43.2	$\mathbf{-657.4 \mp 12.3}$	
DBLP-link	-1923.9 ∓ 19.4	-2124.6 ∓ 26.4	$\mathbf{-1049.6 \mp 7.5}$	
Hypertext	-5276.7 ∓ 9.6	-5281.4 ∓ 10.3	$\mathbf{-2923.2 \mp 0.0}$	
Newcomb	-1075.0 ∓ 47.6	-1098.1 ∓ 48.0	-876.7 ∓ 0.0	
Freeman	-658.5 ∓ 19.6	-664.1 ∓ 19.2	$\mathbf{-405.2 \mp 0.0}$	
Coleman	-1500.8 ∓ 63.7	-1532.8 ∓ 64.2	-1003.9 ∓ 0.0	
Dataset	MMSB	IRM	LFRM	DRIFT
Kapferer	-618.4 ∓ 59.8	-658.6 ∓ 70.3	-865.1 ∓ 70.1	-783.2 ∓ 92.3
Sampson	-353.0 ∓ 16.3	-366.8 ∓ 0.6	-332.2 ∓ 16.9	-275.2 ∓ 52.0
Student-net	-881.4 ∓ 29.9	-1201.2 ∓ 1.6	-1069.6 ∓ 42.2	-905.8 ∓ 46.3
Enron	-1512.5 ∓ 6.5	-2264.8 ∓ 26.2	-1742.9 ∓ 36.0	-1492.3 ∓ 13.2
Senator	-713.2 ∓ 64.2	-843.6 ∓ 23.5	-673.2 ∓ 43.6	-678.6 ∓ 48.5
DBLP-link	-2082.0 ∓ 12.0	-2953.1 ∓ 4.9	-1746.5 ∓ 15.4	-1426.1 ∓ 46.2
Hypertext	-4083.5 ∓ 77.8	-5432.7 ∓ 19.6	-3747.5 ∓ 94.3	-3942.3 ∓ 48.5
Newcomb	-1835.2 ∓ 14.2	-1965.9 ∓ 1.8	-1203.0 ∓ 14.7	$\mathbf{-789.3 \mp 63.2}$
Freeman	-673.5 ∓ 73.9	-728.9 ∓ 66.9	-917.2 ∓ 35.7	-794.2 ∓ 66.2
Coleman	-1302.8 ∓ 130.2	-689.5 ∓ 3.2	-606.7 ∓ 65.1	$\mathbf{-546.1 \mp 26.9}$

Figure 3.7: AUC Performance (95% Confidence Interval = Mean $\mp\times$Standard Deviation)

Dataset	MTV-g	MTV-s	MTI	
Kapferer	0.816 ∓ 0.074	0.816 ∓ 0.011	$\mathbf{0.928 \mp 0.000}$	
Sampson	0.804 ∓ 0.000	0.821 ∓ 0.098	$\mathbf{0.927 \mp 0.000}$	
Student-net	0.867 ∓ 0.030	0.877 ∓ 0.095	0.934 ∓ 0.000	
Enron	0.834 ∓ 0.097	0.853 ∓ 0.143	0.920 ∓ 0.001	
Senator	0.849 ∓ 0.129	0.839 ∓ 0.046	$\mathbf{0.931 \mp 0.001}$	
DBLP-link	0.831 ∓ 0.046	0.816 ∓ 0.017	$\mathbf{0.926 \mp 0.000}$	
Hypertext	0.861 ∓ 0.029	0.843 ∓ 0.027	$\mathbf{0.901 \mp 0.023}$	
Newcomb	0.814 ∓ 0.049	0.795 ∓ 0.090	0.931 ∓ 0.000	
Freeman	0.875 ∓ 0.133	0.862 ∓ 0.041	$\mathbf{0.915 \mp 0.000}$	
Coleman	0.891 ∓ 0.067	0.872 ∓ 0.052	0.928 ∓ 0.000	
Dataset	MMSB	IRM	LFRM	DRIFT
Kapferer	0.893 ∓ 0.001	0.751 ∓ 0.016	0.891 ∓ 0.034	0.905 ∓ 0.013
Sampson	0.836 ∓ 0.002	0.738 ∓ 0.005	0.841 ∓ 0.012	0.855 ∓ 0.029
Student-net	0.938 ∓ 0.001	0.809 ∓ 0.004	0.862 ∓ 0.076	$\mathbf{0.949 \mp 0.015}$
Enron	0.907 ∓ 0.013	0.820 ∓ 0.082	0.894 ∓ 0.073	$\mathbf{0.956 \mp 0.079}$
Senator	0.880 ∓ 0.022	0.829 ∓ 0.064	0.892 ∓ 0.056	0.925 ∓ 0.076
DBLP-link	0.918 ∓ 0.000	0.817 ∓ 0.010	0.891 ∓ 0.062	0.891 ∓ 0.034
Hypertext	0.844 ∓ 0.008	0.788 ∓ 0.015	0.853 ∓ 0.042	0.871 ∓ 0.010
Newcomb	0.836 ∓ 0.001	0.765 ∓ 0.013	0.879 ∓ 0.041	$\mathbf{0.960 \mp 0.027}$
Freeman	0.867 ∓ 0.001	0.790 ∓ 0.008	0.883 ∓ 0.026	0.897 ∓ 0.022
Coleman	0.928 ∓ 0.001	0.888 ∓ 0.004	0.929 ∓ 0.018	$\mathbf{0.945 \mp 0.052}$

Larger data size results

We also conduct the experiments with a larger synthetic dataset ($N = 100, T = 20$). With the construction the same as previously, we increase the role number to 5 and set the role-compatibility matrix as in Figure 3.8.

$$\begin{vmatrix} 0.95 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.1 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.05 & 0.1 & 0.9 & 0.05 & 0 \\ 0.05 & 0 & 0.05 & 0.9 & 0.15 \\ 0 & 0.05 & 0.1 & 0.05 & 0.9 \end{vmatrix}$$

Figure 3.8: Larger dataset's role-compatibility matrix

We set 5 groups in this network, with the group sizes as $[35, 20, 20, 20, 5]$ and the mixed-membership distributions for each of the groups as $[0.8, 0.1, 0, 0.05, 0.05; 0.02, 0.85, 0.05, 0.03, 0.05; 0.1, 0, 0.9, 0, 0; 0.05, 0.1, 0, 0.85, 0; 0, 0.2, 0, 0.4, 0.4]$. The detailed results are also given in Table 3.3. As we can see, our MTI model still achieves the best performance of all the models.

3.4.2 Real World Datasets Performance

We select 10 real world datasets for benchmark testing: Kapferer (Kapferer 1972), Sampson (Sampson 1969)(Breiger, Boorman & Arabie 1975), Student-net, Enron (Klimt & Yang 2004), Senator(Ho et al. 2011), DBLPlink (Asur, Parthasarathy & Ucar 2009)(Lin, Chi, Zhu, Sundaram & Tseng 2009), Hypertext (Isella, Stehl, Barrat, Cattuto, Pinton & Van den Broeck 2011), Newcomb (Newcomb 1961), Freeman (Freeman & Freeman 1979), Coleman (Coleman et al. 1964). Their detailed information, including the number of nodes, the number of edges, edge types and time intervals, are given in Table 3.5. Following a general test on the training log-likelihood of the training data and AUC (Area Under the ROC Curve) of the test data, we give a more detailed elaboration on three selected datasets below.

Table 3.5: Dataset Information

Dataset	Nodes	Edge	Time	Link Type
Kapferer	39	256	2	friends
Sampson	18	168	3	like
Student-net	50	351	3	friends
Enron	151	1980	12	email
Senator	100	5786	8	vote
DBLPlink	100	5706	10	citation
Hypertext	113	7264	10	contact
Newcomb	17	1020	15	contact
Freeman	32	357	2	friends
Coleman	73	506	2	co-work

General performance on Training log-likelihood and AUC value

We use a 5-fold cross validation method to certify our model's performance on the real world datasets. The hyper-parameters γ, κ, α are sampled according to the sampling strategy mentioned in Section 3.3. Each experiment is run 10 times and we report their mean and standard deviation in Table 3.6 and Table 3.7.

In these two tables, the black bold type denotes the best value in each row. As we can see, our MTI model achieves the best values in 8 of the 10 datsets on the training log-likelihood and 6 of the 10 datasets on the AUC value. In the remaining datasets, while our MTI model's performance is still quite competitive, the DRIFT model has the best values, possibly due to that in these datasets, all associated communities from both nodes are considered in generating the link between these two nodes (Miller et al. 2009). The MTV models still do not perform well enough, for the reason previously given. The IRM's results are the worst, which reflects the fact that the simple structure (i.e., each node occupies only one class) may not be enough to capture the full structure in relational learning.

3.4.3 Kapferer Tailor Shop

The Kapferer Tailor Shop data (Nowicki & Snijders 2001) records interactions in a tailor shop at two time points. In this time period, the employees in the shop are negotiating for higher wages. The dataset is of particular interest because two strikes occur after each time point, with the first failing and the second succeeding.

We mainly use the "work-assistance" interaction matrix in the dataset. The employees have 8 occupations: head tailor (19), cutter (16), line 1 tailor (1-3, 5-7, 9, 11-14, 21, 24), button machiner (25-26), line 3 tailor (8, 15, 20, 22-23, 27-28), Ironer (29, 33, 39), cotton boy (30-32, 34-38) and line 2 tailor (4, 10, 17-18).

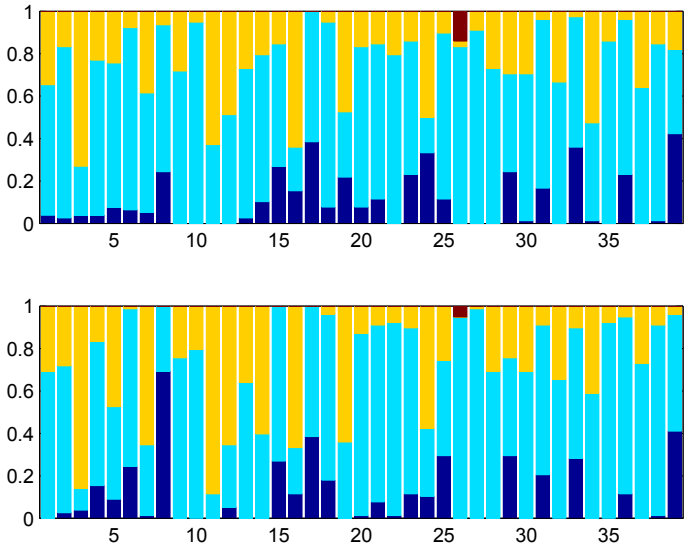

Figure 3.9: The MTI model's Performance on Kapferer Tailor Shop Dataset.

Figure 3.9 shows the mixed-membership distribution's evolving behaviour. The x axis stands for the nodes, while the y axis represents the mixed-membership distribution. Different colors are interpreted as the communities we have discovered. The top barchart describes all the employees' mixed-membership distribution in Time 1, while the bottom one illustrates the one

in Time 2. In Figure 3.9, we can see that the yellow communities at Time 2 are larger than those at Time 1, which means that people tend to have another community at Time 2, rather than being mostly dominated by one large group at Time 1. This larger "yellow" community may be the result of the first failed strike, after which employees start to shift to the "minor" (yellow) community for a successful strike.

3.4.4 Sampson Monastery Dataset

The Sampson Monastery dataset is used here to extend the study. There are 18 monks in this dataset, and their social linkage data is collected at 3 different time points with various interactions. Here, we especially focus on the like-specification. In the like-specification data, each monk selects three monks as his closest friends. In our settings, we mark the selected interactions as 1, otherwise 0. Thus, an $18 \times 18 \times 3$ social network dataset is constructed, with each row having three elements valued at 1.

According to previous studies (Kim et al. 2012)(Xing et al. 2010), the monks are divided into 4 communities: Young Turks, Loyal Opposition, Outcasts and an interstitial group.

Figure 3.10 shows the detail results of the MTI model. As three communities have been detected, we put all the results in a 2-simplex, in which we denote the communities as A, B and C. For trajectory convenience, we also color the nodes according to the special group to which belong. As we can see, these groups' behavior differs significantly in the figure. The Loyal Opposition group lies close to C, and the interstitial group tends to belong to A. Both of their mixed-membership distribution is stable across time. On the Outcasts and Young Turks groups, they lie much closer to B.

We also provide the role-compatibility matrix in Figure 3.11 for comparison. Compared to the result in (Xing et al. 2010), our results have a larger compatibility value within the same role. Also, the first role's value in our model is 0 while it is about 0.6 in (Xing et al. 2010).

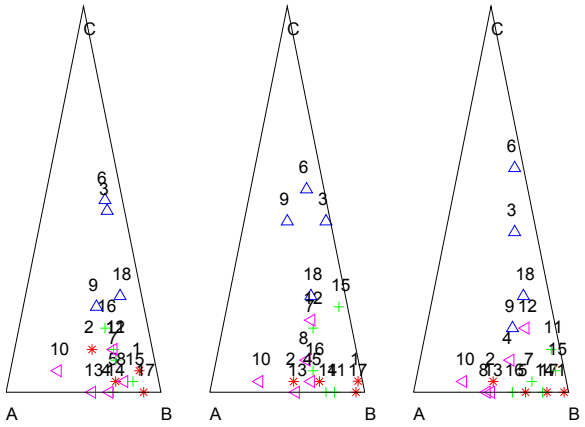

Figure 3.10: The nodes' mixed-membership distribution of the MTI model on Sampson Monastery Dataset (from Left to Right: Time 1-3.) Blue, Loyal Opposition; Red, Outcasts; Green, Young Turks; Magenta, interstitial group.

$$\begin{vmatrix} 0.09 & 0 & 0.0 \\ 0.05 & 0.99 & 0.02 \\ 0.01 & 0 & 0.96 \end{vmatrix} \qquad \begin{vmatrix} 0.01 & 0 & 0.03 \\ 0.02 & 0.78 & 0 \\ 0.02 & 0 & 0.67 \end{vmatrix}$$

Figure 3.11: Role Compatibility Matrix (Left: MTV-g; Right: MTI)

3.4.5 Hypertext 2009 dynamic contact network

This dataset (Isella et al. 2011) is collected from the ACM Hypertext 2009 conference. 113 conference attendees volunteered to wear radio badges that recorded their face-to-face contact during the conference. The original data is composed of the records such as "(t, i, j)", where t is the communication time and i, j are the attendees' ID. By adaptively cutting the whole time period into 10 parts and noting the interaction data as "1" if communicated during the time stamps, we obtain a $113 \times 113 \times 10$ binary matrix. Figure 3.12 displays the dynamic behavior of the nodes' mixed-membership distributions and the corresponding role-compatibility matrix. The numbers on

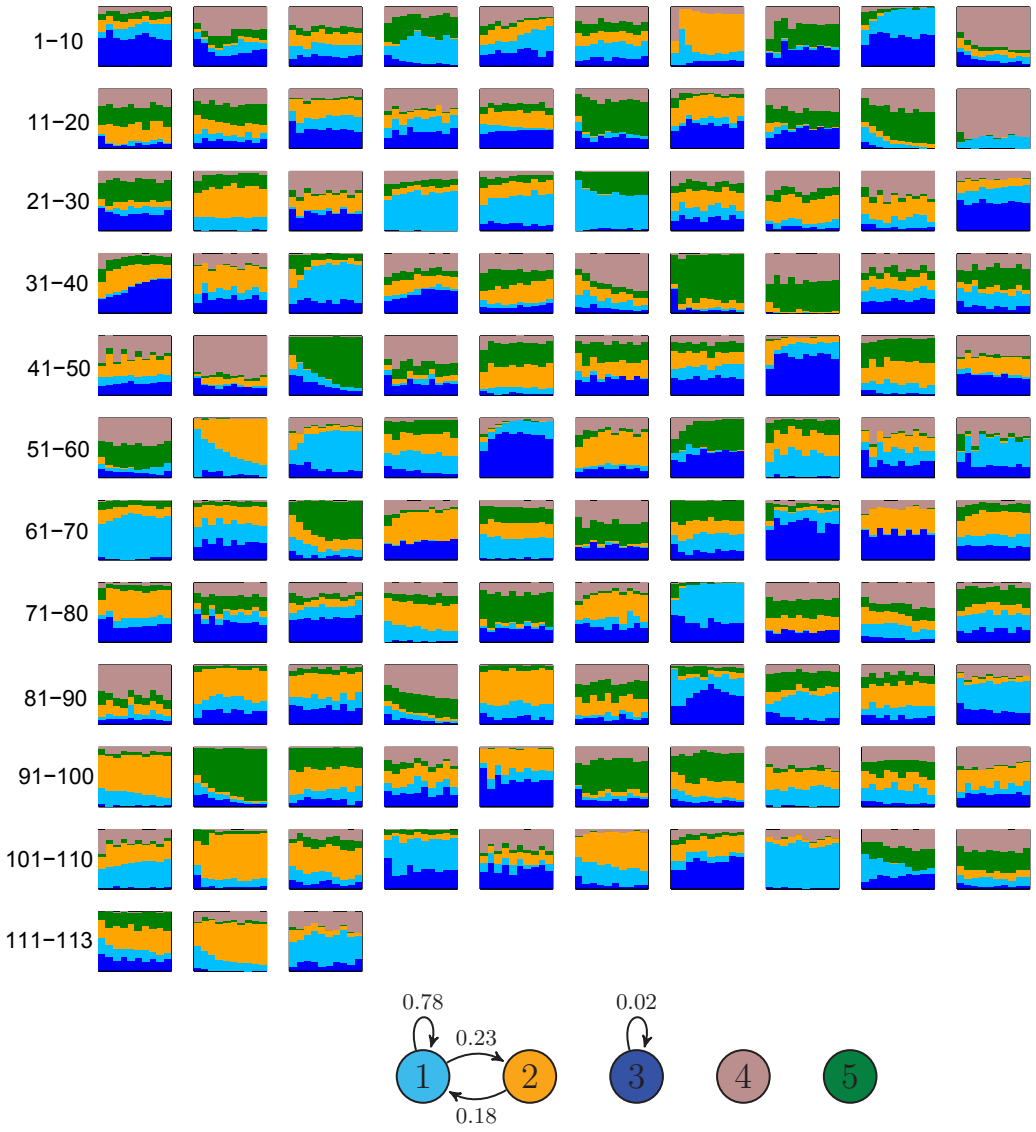

Figure 3.12: The MTI model's performance on the Hypertext 2009 dynamic contact network.

the left side denote the orders of nodes. Each bar chart represents the dynamic behavior of one node's mixed-membership distribution, where the x axis stands for the 10 time stamps. Different colors are interpreted as the communities we have discovered, and their role-compatibility is represented below the bar chart.

As we can see, almost half of all the mixed-membership distributions fluctuated during these time stamps. This phenomenon coincides with our common knowledge that people at academic conferences tend to communicate causally. Thus, people's roles may change during different time stamps.

The value of the role-compatibility matrix we learned is focused on the "sky blue" community, whose intra-role compatibility value is 0.6932. It has a small probability of interaction with other communities. The other community's compatibility value is almost 0. This might be the reason for sparsity in the interaction data.

Here we would specially mention node 108. In the record, this person is always the first to communicate with others on each of the three days. His mixed-membership distribution is mainly composed of the "sky blue" community 1, which indicates he could be an organizer of this conference. The other nodes with mixed-membership distribution dominated by community 1, such as nodes $24, 53, 61$, all play actively with others according to the record.

Another interesting phenomenon is that the nodes contain the "orange" community 2 will interact with community 1 with a probability of 0.2. This might be an indication that most of the attendees will communicate with the organizers for various reasons.

3.5 Summary

Modeling complex networking behaviors in a dynamic setting is crucial for widespread applications such as social media, social networks, online business and market dynamic analysis. This challenges existing learning systems,

which have limited power to address the dynamics. In this Chapter, we have provided a generalised and flexible framework for describing the coupling relations between the mixed-membership distributions over the time, which is to further improve the popular mixed-membership stochastic blockmodel by allowing a network to have infinite types of communities with relationships that change across time. Through two different strategy of treating the time-dependent coupling relations, we have shown our models' effect on depicting the mixed-membership distributions, as well as the communities' time evolving behaviour. Further, both Gibbs sampling and adapted Slice-Efficient sampling have been used to inference the desired target distribution. The quantity analysis on the MCMC's convergence behavior, including the convergence test, autocorrelation function, etc., have been provided to enhance the inference performance.

3.6 Limitation & future work

We should also note that our current DIM3 model only captures the incremental changes in the social network. The dynamic changes of hidden membership indicators s_{ij}^t, r_{ij}^t only reflect the changes of individual people. While there exist other kinds of dynamic changes in the social network, such as the whole group's re-organization or a subset of groups' changing, our model need to be adjusted on focusing the whole groups' dynamic behaviour, such as the compatibilities' change. Also, other methods, such as the change detection methods, might be an alternative candidature to address these issues.

On describing the coupling relations between the mixed-membership distributions, the Dependent Dirichlet Process (DDP) (MacEachern 1999) provides an alternative. Among the various constructions of the DDP (Caron, Davy & Doucet 2007)(Chung & Dunson 2011)(Dunson 2006)(Bouguila & Ziou 2010), (Rao & Teh 2009) constructs the DDP by projecting the Gamma Process into different subspaces and normalizing them individually, through which the overlap spaces reflect the correlation. (Lin, Grimson & Fisher 2010)

discusses the intrinsic relationship between the Poisson Process, Gamma Process and Dirichlet Process and uses three operations: superposition, subsampling and point transition to evolve from one Dirichlet process to another, with an elegant and solid theory support. Subsequent work including (Lin & Fisher 2012)(Chen, Ding & Buntine 2012)(Chen, Rao, Buntine & Teh 2013) extends this work from different perspectives.

Apart from this coupling relation's construction, recent developments (Gopalan, Gerrish, Freedman, Blei & Mimno 2012)(Yin, Ho & Xing 2013) in large-scale learning of latent space modelling gives many approaches to the future works. These attractive improvements include a parsimonious link modelling (Gopalan et al. 2012), which reduces the parameter size from $\mathcal{O}(n^2K^2)$ to $\mathcal{O}(n^2K)$; the utilization of stochastic variational inference method (Hoffman, Blei, Wang & Paisley 2013); a triangular representation of network (Hunter, Goodreau & Handcock 2008)(Yin et al. 2013), which could reduce the parameter size to $\mathcal{O}(nK^2)$. Through these new findings, it is hoped to scale the models to millions of nodes and hundreds of communities.

Chapter 4

Copula Mixed-Membership Stochastic Blockmodel for Intra-group Correlation

4.1 Introduction

Community modeling is an important but challenging topic which has seen applications in various settings including social-media recommendation (Tang & Liu 2010), customer partitioning, discovering social networks, and partitioning protein-protein interaction networks (Girvan & Newman 2002)(Fortunato 2010). Quite a few models have been proposed in the last few years to address these problems; some earlier examples include stochastic blockmodel (Nowicki & Snijders 2001), and its infinite community case - infinite relational model (IRM) (Kemp et al. 2006), both assume that each node has one latent variable to directly indicate its community membership, dictated by a single distribution of communities. Their aim is to partition a network of nodes into different communities based on the pair-wise, directional binary observations.

A typical need and challenge in community modeling is to capture the complex interactions amongst the nodes in different applications. According-

ly, several variants of IRM were proposed, including the mixed membership stochastic blockmodel (MMSB) (Airoldi et al. 2008), in which multiple roles (membership indicators) can possibly be played by one node. Each node has its own "membership distribution", and its relation with all other nodes is generated from it. For any two nodes, having determined their corresponding membership indicator pair, their (directional) interactions are generated from a so-called, "role-compatibility matrix" with its row and column indexed by this pair. One mentionable development of MMSB is the nonparametric metadata dependent relational model (NMDR) (Kim et al. 2012), which modifies MMSB by incorporating each node's metadata information into the membership distribution.

However, all of the MMSB-typed models make the assumption that, for each relation between two nodes, their corresponding membership indicator pairs were determined independently. This may limit the way membership indicators can be distributed. In fact, under many social network settings, certain known group members may have higher correlated interactions towards the ones within the same group. For instance, in a company, IT support team members tend to co-interact with each other more than with employees of other departments. Another example is that teenagers may have similar "likes" or "dislikes" on certain topics, compared with the views they may hold towards people of other age groups. MMSB-typed models overlook such interactions within a group and thus cannot fully capture the intrinsic interactions within a network.

In reality, within a social networking context, it is important to incorporate group member interactions (here called intra-group correlations) into the modeling of membership indicators. Especially, the strong coupling relations inside each community plays a critical role in forming the interactions between the nodes. While some communities may promote cohesive interactions, others may produce repelled interactions. After introducing these intra-group correlations (which is also named as **Synchronous coupling and Exclusive coupling**), it is important that at the same time, we do not

alter membership indicators' distributions themselves, so that their interactions to people outside of the known subgroups are unaffected.

Accordingly, in this chapter, a Copula function (Nelsen 2006)(McNeil & Nešlehová 2009) is introduced to MMSB, forming a copula mixed-membership stochastic blockmodel (cMMSB), for modeling the intra-group correlations. With cMMSB, we can flexibly apply various Copula functions towards different subsets of pairs of nodes while maintaining the original marginal distribution of each of the membership indicators. We develop ways in which a bivariate Copula can be used for two distributions of indicators, enjoying infinitely possible values. Under the framework, we can incorporate different choices of Copula functions to suit the need of the applications. With different Copula functions imposed on the different groups of nodes, each of the Copula function's parameters will be updated in accordance with the data. What is more, we also give two analytical solutions to calculate the conditional marginal density to the two indicator variables, which plays a crucial role in our likelihood calculation and also creates a new way of calculating a deterministic relationship between multiple variables in a graphical model.

The rest of the chapter is organized as follows. In Section 4.2, we give a literature review on relational models and a brief overview of the Copula model. In Section 4.3, we present the main model, especially the details of our Copula-based MMSB. We further provide two "collapsed" sampling methods for the conditional probabilities, described in Section 4.4. In Section 4.5, we show the experimental results of our model, using both the synthetic and real-world social network data. In Section 4.6, we would briefly conclude this chapter.

4.2 Copula Model

Here we describe very briefly a bivariate copula function $C(u, v)$, which is a Cumulative Distribution Function over the interval $[0, 1] \times [0, 1]$ with the uniform marginal distribution (Nelsen 2006). This correlation representation

is extremely useful since we have the following theorem:

Theorem 4.1 *Sklar's Theorem: Let X and Y be random variables with distribution functions F and G respectively and joint distribution function H. Then there exists a Copula C such that for all $(x, y) \in R \times R$:*

$$H(x, y) = C(F(x), G(y)) \tag{4.1}$$

C is unique if F and G are continuous, then the joint probability density function is:

$$h(x, y) = c(F(x), G(y)) \cdot f(x)g(y) \tag{4.2}$$

Here $c(u, v) = \partial^2 C(u, v)/\partial u \partial v$ is noted for the copula density function.

Sklar's theorem ensures the uniqueness of copula function $C(F(x), G(y))$ once the joint distribution $h(x, y)$ and its two marginal distributions $f(x)$ and $g(y)$ are known. The modification of a Copula function does not change the marginal distributions, which serves the purpose of this chapter.

The popularity of copula models from various applications also meant the availability of different choices of copula functions to suit various applications. The commonly used copula function includes Gaussian Copula (Gaussian, t), Archimedean Copula (Clayton, Gumbel, Frank, etc.), and Empirical Copula. For a comprehensive survey of copula functions, please refer to (Nelsen 2006).

4.3 Graphical Model Description

The generative process of graphical modeling is illustrated below:

C1: $\boldsymbol{\beta} \sim GEM(\gamma)$

C2: $\{\pi_i\}_{i=1}^{n} \sim DP(\alpha \cdot \boldsymbol{\beta})$

C3: $\begin{cases} (u_{ij}, v_{ij}) \sim Copula(\theta), & g_{ij} = 1, (i,j) \text{belongs to the sub-group of interest;} \\ u_{ij}, v_{ij} \sim U(0,1), & g_{ij} = 0, (i,j) \text{is under the traditional MMSB framework.} \end{cases}$

C4: $s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij})$

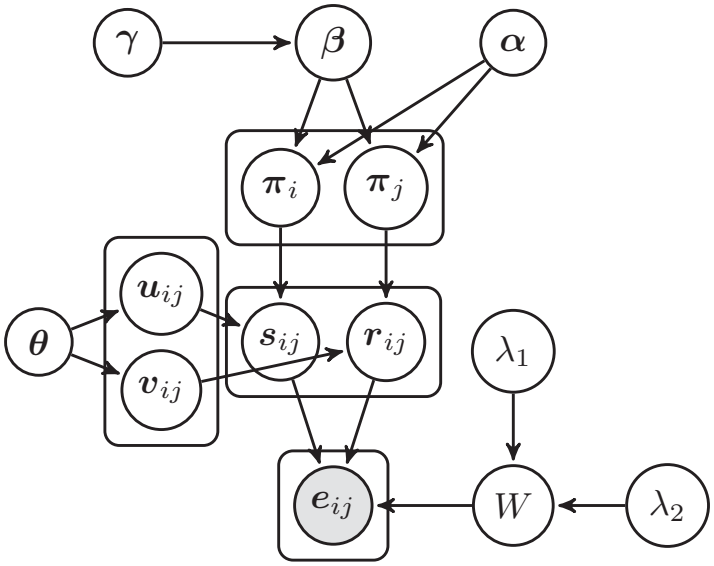

Figure 4.1: Graphical model of Copula MMSB

C5: $W_{k,l} \sim Beta(\lambda_1, \lambda_2), \forall k, l$;

C6: $\boldsymbol{e}_{ij} \sim Bernoulli(W_{s_{ij}, r_{ij}})$.

Here $g_{ij} = 1$ in C3 denotes that the node pair (i, j) belongs to the sub-group of interest, i.e., s_{ij}, r_{ij} are implicitly correlated, while $g_{ij} = 0$ means (s_{ij}, r_{ij}) are modelled using traditional MMSB. In C4, $\Pi_i^{-1}(u_{ij}) = \{\min k : \sum_{q=1}^{k} \pi_{iq} \geq u_{ij}\}$ denotes the interval of π_i that u_{ij} belongs into, and similar notation is applied to $\Pi_j^{-1}(v_{ij}) = \{\min k : \sum_{q=1}^{k} \pi_{jq} \geq v_{ij}\}$.

For a simplified illustration, we divide the generative model into three sub-models: (1) "mixed membership distribution modelling", (2) "copula incorporated membership indicator pair" and (3) "binary observation modelling", with their details elaborated in the following sections.

Mixed Membership Distribution Modeling

C1-C2 are for the generation of each node's mixed membership distribution. The number of communities, i.e., k is an important factor in mixed membership distribution models. Therefore, we consider two possibilities here. The

first is to use a fixed k. As the graphical model in Fig. 4.1 shows, for all the mixed-membership distributions $\{\pi_i\}_{i=1}^n$, there is a common parent node $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ typically has a "non-informative" symmetric Dirichlet prior, i.e., $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \sim Dir(\gamma, \ldots, \gamma)$ (Airoldi et al. 2008). The appropriate choice of k is determined by the model selection method, such as the BIC criterion (Schwarz 1978), which is commonly used in (Airoldi et al. 2008)(Xing et al. 2010).

The second solution is applicable for the uncertain number of communities, which is often the case under many social network settings. The usual approach is to use the Hierarchical Dirichlet Process (HDP) (Teh et al. 2006) prior with $\boldsymbol{\beta}$ distributed from a GEM(γ), i.e., $\boldsymbol{\beta}$ is obtained via a stick-breaking construction (Sethuraman 1994) with each of its components $\boldsymbol{\beta}_k = u_k \prod_{l=1}^{k-1}(1-u_l), u_l \sim Beta(1, \gamma)$.

After obtaining their parent's node $\boldsymbol{\beta}$, we can sample our mixed-membership distribution $\{\pi_i\}$ independently from (Airoldi et al. 2008)(Koutsourelakis & Eliassi-Rad 2008): $\pi_i \sim \begin{cases} Dir(\alpha \cdot \boldsymbol{\beta}), & \text{fixed } k; \\ DP(\alpha \cdot \boldsymbol{\beta}), & \text{uncertain } k. \end{cases}$ For the notational clarity, we concentrate our discussion on the uncertain k case without delicately mentioning its finite counterpart, as the finite k case can be trivially derived.

Copula Incorporated Membership Indicator Pair

Our main work of c-MMSB is displayed in phases C3-C4. We consider two cases in this chapter for the intra-group correlation modeling: full correlation and partial correlation.

Full correlation: i.e., intra-group correlation for all the nodes. We assume each pair of nodes, i.e., all relations of the entire population are using the same Copula function. As we will see in the experimental section that, flexible modelling can still be achieved under this assumption, as parameters of a Copula can vary to support various form of relations.

Partial correlation: i.e., intra-group correlation are applied to only a subset of the nodes. With a definition of one subgroup, we use the Copula function on this specific subgroup and the others remain unchanged.

For traditional MMSB, the corresponding membership indicators within one pair (s_{ij}, r_{ij}) are independently sampled from their membership distributions, i.e., $s_{ij} \sim \pi_i, r_{ij} \sim \pi_j$. Using the definition of $\{\Pi_i^{-1}(\cdot)\}_{i=1}^n$ from Section 4.3, this is equivalently expressed as:

$$\begin{aligned} u_{ij} \sim U(0,1), v_{ij} \sim U(0,1); \\ s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij}). \end{aligned} \tag{4.3}$$

As discussed in the introduction, we are motivated by examples within social network settings, in which membership indicators from a node may well be correlated with other membership indicators in an intra-group point of view. People's interactions with each other within the group may more likely (or less likely) belong to the same category, i.e., (s_{ij}, r_{ij}) has higher (or lower) density in some regions of the discrete space $(1, 2, \ldots, \infty)^2$, which may not be well described by using only the two independent marginal distributions.

We propose a general framework by employing a Copula function to depict the correlation within the membership indicator pair. This is accomplished by the joint sampling of uniform variables (u_{ij}, v_{ij}) (in Eq. (4.3).) from the Copula function, instead of from two independent uniform distributions. More precisely, the membership indicator pair is obtained using:

$$\begin{aligned} \forall g_{ij} = 1 : (u_{ij}, v_{ij}) \sim Copula(u, v|\theta); \\ s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij}). \end{aligned} \tag{4.4}$$

Using various Copula priors over the pair (u_{ij}, v_{ij}), we are able to more appropriately express the way in which the membership indicator pair $\{s_{ij}, r_{ij}\}$ is distributed, given the different scenarios we are facing. Taking the Gumbel Copula (with larger parameter values) (Nelsen 2006) as an instance, for certain membership indicator pairs ($g_{ij} = 1$), it generates (u_{ij}, v_{ij}) values that more likely have positive correlation, i.e., within $[0, 1]^2$ space, which promotes $s_{ij} = r_{ij}$. Also, the Gaussian Copula ($\theta = -1$) encourages the (s_{ij}, r_{ij}) pair to be different.

Binary Observation Modeling

C5-C6 model the binary observation, which directly follows the previous work (Nowicki & Snijders 2001)(Kemp et al. 2006) etc. Due to the beta-bernoulli conjugacy, W can be marginalized out and the likelihood of binary observation becomes as follows:

$$\Pr(\boldsymbol{e}|z,\lambda_1,\lambda_2) = \prod_{k,l} \frac{\text{beta}(m^1_{k,l}+\lambda_1, m^0_{k,l}+\lambda_2)}{beta(\lambda_1,\lambda_2)} \tag{4.5}$$

here $z = \{s_{ij}, r_{ij}\}_{i,j=1,\cdots,n}$ denotes all the hidden labels, $\text{beta}(\lambda_1,\lambda_2)$ denotes the beta function with parameters λ_1 and λ_2, $m^1_{k,l}$ is number of link 1 from community k to l, i.e. $m^1_{k,l} = \sum_{s_{ij}=k,r_{ij}=l} e_{ij}$, $m^0_{k,l}$ is number of link 0 from community k to l, i.e. $m^0_{k,l} = \sum_{s_{ij}=k,r_{ij}=l}(1-e_{ij})$.

4.4 Inference & Further Discussion

Let K be the discovered number of communities, a formal and concise representation of Eq. (4.4), i.e. the probability of (s_{ij}, r_{ij}), is:

$$\begin{aligned}\Pr(s_{ij}, r_{ij}) = &\int_{\sum_{d=1}^{K+1}\pi_{jd}=1}\int_{\sum_{d=1}^{K+1}\pi_{id}=1}\int_{(u_{ij},v_{ij})} \\ &\cdot \mathbf{1}\left(s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij})\right) \\ &\cdot dC(u_{ij},v_{ij})dF(\pi_{i1},\cdots,\pi_{iK+1})dF(\pi_{j1},\cdots,\pi_{jK+1})\end{aligned} \tag{4.6}$$

Unfortunately, we cannot bring $\Pr(s_{ij}, r_{ij})$ to an analytical form without any integrals present. However, with some mathematical design, we found that, conditioning on the explicit sample of either (u_{ij}, v_{ij}) or (π_i, π_j), it is possible to obtain a marginalised conditional density in which s_{ij}, r_{ij} is conditioned on either (u_{ij}, v_{ij}) or (π_i, π_j), but not both. Additionally, having a set of variables "collapsed" from the Gibbs sampling, it results in a faster mixing on Markov chains (Liu 1994). Therefore, two corresponding inference schemes are needed. We present both inference below, and name them Marginal conditional on π only method and the Marginal conditional on u, v only respectively:

4.4.1 Marginal Conditional on π only: cMMSB$^{\pi}$

In the Marginal conditional on π only (cMMSB$^{\pi}$ for short) method, the variables of interest include $\{\pi_i\}, \{s_{ij}, r_{ij}\}, \boldsymbol{\beta}$. As mentioned before, we describe the formulation using the infinite communities (uncertain k) case only, its counterpart in the finite communities (fixed k) case can be trivially derived.

Sampling π_i

When a Copula is introduced, $p(\pi_i)$ and $\Pr(s_{ij}|\pi_i)$ are no longer a conjugate pair. Therefore, we resort to the use of Metropolis-Hastings (M-H) Sampling in each (τ)-th MCMC iteration.

For each node i, π_i's posterior distribution is formed as Eq. (4.7), where $p_{ij}^{s_{ij}r_{ij}}(\pi_i, \pi_j)$ is defined in Eq. (4).

$$\begin{aligned} & p(\pi_i|\alpha, \boldsymbol{\beta}, \{s_{ij}, r_{ij}\}_{i,j}) \\ \propto & \prod_{k=1}^{K+1} \pi_{ik}^{\alpha\boldsymbol{\beta}_k - 1} \cdot \prod_{j=1}^{n} \left[p_{ij}^{s_{ij}r_{ij}}(\pi_i, \pi_j) p_{ji}^{s_{ji}r_{ji}}(\pi_j, \pi_i)\right] \end{aligned} \tag{4.7}$$

The Corresponding proposal distribution of π_i for the above M-H is a posterior Dirichlet distribution in the form of (i.e., π_i's posterior distribution under the MMSB framework):

$$q(\pi_i^*|\alpha, \boldsymbol{\beta}, \{s_{ij}, r_{ij}\}_{i,j}) \propto \prod_{k=1}^{K+1} [\pi_{ik}^*]^{\alpha\boldsymbol{\beta}_k + N_{ik} - 1} \tag{4.8}$$

Then the acceptance ratio becomes:

$$A(\pi_i^*, \pi_i^{(\tau)}) = \min(1, a) \tag{4.9}$$

$$a = \frac{\prod_{j=1}^{n} \left[p_{ij}^{s_{ij}r_{ij}}(\pi_i^*, \pi_j) p_{ji}^{s_{ji}r_{ji}}(\pi_j, \pi_i^*)\right]}{\prod_{j=1}^{n} \left[p_{ij}^{s_{ij}r_{ij}}(\pi_i^{(\tau)}, \pi_j) p_{ji}^{s_{ji}r_{ji}}(\pi_j, \pi_i^{(\tau)})\right]} \cdot \frac{\prod_{k=1}^{K+1} \left[\pi_{ik}^{(\tau)}\right]^{N_{ik}}}{\prod_{k=1}^{K+1} [\pi_{ik}^*]^{N_{ik}}} \tag{4.10}$$

Sampling s_{ij} and r_{ij}

As e_{ij} is dependent on both $\{s_{ij}, r_{ij}\}$, a joint sampling of $\{s_{ij}, r_{ij}\}$ is implemented as:

$$\begin{aligned} & \Pr(s_{ij}, r_{ij} | e_{ij}, \lambda_1, \lambda_2, \theta, \pi_i, \pi_j, m_{s_{ij}, r_{ij}}^{-e_{ij}}) \\ \propto \quad & \Pr(s_{ij}, r_{ij} | \pi_i, \pi_j, \theta) \cdot \Pr(e_{ij} | s_{ij}, r_{ij}, \lambda_1, \lambda_2, m_{s_{ij}, r_{ij}}^{-e_{ij}}) \end{aligned} \tag{4.11}$$

Where (w.l.o.g., we assume $s_{ij} = k, r_{ij} = l$)

$$\begin{aligned} & \Pr(e_{ij} | s_{ij} = k, r_{ij} = l, \lambda_1, \lambda_2, m_{k,l}^{-e_{ij}}) \\ \propto & P(e_{ij}, \boldsymbol{e} \backslash \{e_{ij}\}, s_{ij} = k, r_{ij} = l, , m_{k,l}^{-e_{ij}}, \lambda_1, \lambda_2) \\ = & \int_{W_{k,l}} P(e_{ij} | s_{ij} = k, r_{ij} = l, W_{k,l}) \cdot p(W_{k,l} | \lambda_1, \lambda_2) \\ & \cdot \prod_{s_{i'j'} = k, r_{i'j'} = l, i'j' \neq ij} P(e_{i'j'} | W_{k,l}) dW_{k,l} \\ = & \int_{W_{k,l}} W_{k,l}^{e_{ij} + m_{k,l}^1 + \lambda_1 - 1} (1 - W_{k,l})^{1 - e_{ij} + m_{k,l}^0 + \lambda_2 - 1} dW_{k,l} \\ = & \frac{\Gamma(e_{ij} + m_{k,l}^1 + \lambda_1) \Gamma(1 - e_{ij} + m_{k,l}^0 + \lambda_2)}{\Gamma(1 + m_{k,l} + \lambda_1 + \lambda_2)} \end{aligned} \tag{4.12}$$

Here $m_{k,l} = \sum_{i'j'} \mathbf{1}(s_{i'j'} = k, r_{i'j'} = l)$, $m_{k,l}^1 = \sum_{s_{i'j'} = k, r_{i'j'} = l} e_{i'j'}$, and $m_{k,l}^0 = m_{k,l} - m_{k,l}^1$. Thus, we obtain

$$\Pr(e_{ij} | s_{ij}, r_{ij}, \lambda_1, \lambda_2, m_{s_{ij}, r_{ij}}^{-e_{ij}}) = \begin{cases} m_{s_{ij}, r_{ij}}^{1, -e_{ij}} + \lambda_1, & e_{ij} = 1; \\ m_{s_{ij}, r_{ij}}^{0, -e_{ij}} + \lambda_2, & e_{ij} = 0. \end{cases} \tag{4.13}$$

On the first term of the r.h.s. in Eq. (4.11), we define $p_{ij}^{kl}(\pi_i, \pi_j) \equiv \Pr(s_{ij} = k, r_{ij} = l | \pi_i, \pi_j, \theta), \forall g_{ij} = 1$, and let $C(u_{ij}, v_{ij} | \theta)$ be the chosen Copula cumulative distribution function (c.d.f.) with parameter θ. Given the explicit values of π_i, π_j, we can integrate over all u_{ij}, v_{ij} to compute the probability mass of the indicator pair $(s_{ij} = k, r_{ij} = l), k, l \in \{1, \cdots, K+1\}$:

$$\begin{aligned} p_{ij}^{kl}(\pi_i, \pi_j) = & \int_{\hat{\pi}_i^{k-1}}^{\hat{\pi}_i^k} \int_{\hat{\pi}_j^{l-1}}^{\hat{\pi}_j^l} dC(u, v | \theta) \\ = & C(\hat{\pi}_i^k, \hat{\pi}_j^l) + C(\hat{\pi}_i^{k-1}, \hat{\pi}_j^{l-1}) - C(\hat{\pi}_i^k, \hat{\pi}_j^{l-1}) - C(\hat{\pi}_i^{k-1}, \hat{\pi}_j^l) \end{aligned} \tag{4.14}$$

Here $\hat{\pi}_i^k = \begin{cases} 0, & k = 0; \\ \sum_{q=1}^{k} \pi_{iq}, & k > 0 \end{cases}$.

Since $\{\pi_i\}_{i=1}^n$ are piecewise functions, we can easily calculate the probability mass in this "rectangular" area. In other cases of $\{g_{ij} = 0\}$, i.e., interaction data e_{ij} falls outside of the correlated relation group, we have $p_{ij}^{kl}(\pi_i, \pi_j) = \pi_{ik}\pi_{jl}$.

It is noted that, using the properties of a Copula function, the marginal distributions of $\Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j, \theta)$ remain π_i and π_j respectively, which becomes that of:

$$\begin{aligned} &\sum_{l=1}^{K+1} \Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j, \theta) = \pi_{ik}; \\ &\sum_{k=1}^{K+1} \Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j, \theta) = \pi_{jl}. \end{aligned} \tag{4.15}$$

Sampling β

An obvious choice for the proposal distribution of $\boldsymbol{\beta}$ used in M-H is its prior $p(\boldsymbol{\beta}|\gamma) = GEM(\gamma)$. However, this proposal can be non-informative, which results in a low acceptance rate. We sample $\boldsymbol{\beta}^*$ conditioned on an auxiliary variable $\boldsymbol{m}$: $(\boldsymbol{\beta}_1^*, \cdots, \boldsymbol{\beta}_K^*, \boldsymbol{\beta}_{K+1}^*) \sim Dir(\boldsymbol{m}_1, \cdots, \boldsymbol{m}_K, \gamma)$, in order to increase the M-H's acceptance rate, where $\boldsymbol{m}$ are sampled in accordance with the method proposed in (Teh et al. 2006). However, instead of sampling $\boldsymbol{\beta}$ directly from $\boldsymbol{m}$ as in (Teh et al. 2006), we only use it for our proposal distribution, as we have explicitly sampled $\{\pi_i\}_{i=1}^n$. The acceptance ratio is hence:

$$A(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(\tau)}) = \min(1, a) \tag{4.16}$$

$$a = \frac{\prod_{i=1}^n \left[\prod_{d=1}^{K+1} \Gamma(\alpha\boldsymbol{\beta}_d^{(\tau)}) \cdot \pi_{id}^{\alpha\boldsymbol{\beta}_d^*}\right]}{\prod_{i=1}^n \left[\prod_{d=1}^{K+1} \Gamma(\alpha\boldsymbol{\beta}_d^*) \cdot \pi_{id}^{\alpha\boldsymbol{\beta}_d^{(\tau)}}\right]} \cdot \frac{\prod_{d=1}^{K} \left[\boldsymbol{\beta}_d^{(\tau)}\right]^{\boldsymbol{m}_d - \gamma}}{\prod_{d=1}^{K} [\boldsymbol{\beta}_d^*]^{\boldsymbol{m}_d - \gamma}} \tag{4.17}$$

Sampling Hyper-parameters

Hyper-parameters γ, α are distributed from a vague gamma prior $\mathcal{G}(1,1)$. Since $\log(\gamma|.)$ is log-concave, we use the Adaptive Rejection Sampling method (Rasmussen 1999) to obtain its value. For α, we use the Auxiliary Variable Sampling (Teh et al. 2006) to complete α's update. An M-H Sampling scheme is used to update the Copula function parameter θ. We set the proposal $q(\theta)$ to be the prior of θ, i.e., $\mathcal{G}^*(1,1)$, where $\mathcal{G}^*(x;a,b) = \mathcal{G}(x-1;a,b)\ \forall x > 1$. Then, the acceptance ratio $A(\theta_d^*, \theta_d^{(\tau)})$ becomes that of:

$$A(\theta_d^*, \theta_d^{(\tau)}) = \min(1, a) \tag{4.18}$$

$$a = \frac{\prod_{g_{ij}=1} \left[\Pr(s_{ij}, r_{ij}|\pi_i, \pi_j, \theta_d^*) \Pr(s_{ji}, r_{ji}|\pi_j, \pi_i, \theta_d^*)\right]}{\prod_{g_{ij}=1} \left[\Pr(s_{ij}, r_{ij}|\pi_i, \pi_j, \theta_d^{(\tau)}) \Pr(s_{ji}, r_{ji}|\pi_j, \pi_i, \theta_d^{(\tau)})\right]} \tag{4.19}$$

4.4.2 Marginal Conditional on u and v only: cMMSB^{uv}

In Marginal conditional on u, v only method (cMMSB^{uv} for short), the variables of interest include $\{u_{ij}, v_{ij}\}, \{s_{ij}, r_{ij}\}, \boldsymbol{\beta}$, and an auxiliary variable $\boldsymbol{m}$.

Sampling u_{ij} and v_{ij}

We have used the M-H Sampling for $(u_{ij}, v_{ij}), \forall i, j \in \{1, \ldots, n\}$, due to the nonconjugacy issue. The Copula function is used as its proposal, and therefore, its corresponding acceptance ratio becomes that of:

$$A\left((u_{ij}^{(\tau)}, v_{ij}^{(\tau)}), (u_{ij}^*, v_{ij}^*)\right) = \min(1, a) \tag{4.20}$$

$$\begin{aligned} a = & \frac{I_{u_{ij}^*}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}^*}(h_i^k, \hat{h}_i^k)}{I_{u_{ij}^{(\tau)}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}^{(\tau)}}(h_i^k, \hat{h}_i^k)} \\ & \cdot \frac{I_{v_{ij}^*}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}^*}(h_j^l, \hat{h}_j^l)}{I_{v_{ij}^{(\tau)}}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}^{(\tau)}}(h_j^l, \hat{h}_j^l)} \end{aligned} \tag{4.21}$$

Here $h_i^k, \hat{h}_i^k$'s definitions are the same as in Eq. (7), assuming $s_{ij} = k, r_{ij} = l$.

Sampling s_{ij} and r_{ij}

An alternative "collapsed" sampling method is to integrate over $\{\pi_i\}_{i=1}^n$ while we explicitly sample the values of $\{(u_{ij}, v_{ij})\}_{i,j}$.

Similar as Eq. (4.11), we obtain:

$$\begin{aligned}
&\Pr(s_{ij}=k, r_{ij}=l| \\
&\quad e_{ij}, \lambda_1, \lambda_2, m_{k,l}, u_{ij}, v_{ij}, \{h_i^k\}_k, \{\hat{h}_i^k\}_k, \{h_j^k\}_k, \{\hat{h}_j^k\}_k) \\
&\propto \Pr(s_{ij}=k|u_{ij}, , \{h_i^k\}_k, \{\hat{h}_i^k\}_k) \\
&\quad \cdot \Pr(r_{ij}=l|v_{ij}, \{h_j^k\}_k, \{\hat{h}_j^k\}_k) \cdot \Pr(e_{ij}|\lambda_1, \lambda_2, m_{k,l}) \\
&\propto (I_{u_{ij}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}}(h_i^k, \hat{h}_i^k)) \\
&\quad \cdot (I_{v_{ij}}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}}(h_j^l, \hat{h}_j^l)) \cdot \Pr(e_{ij}|\lambda_1, \lambda_2, m_{k,l})
\end{aligned} \tag{4.22}$$

From Eq. (4.4), given $\{(u_{ij}, v_{ij})\}_{i,j}$'s values, the probabilities $s_{ij}=k$ and $r_{ij}=l$ can be computed independently. The Copula function leaves marginal distributions of s_{ij} and r_{ij} invariant, which remains the same as the classical MMSB, i.e., $\pi_i|\alpha, \beta, \{N_{ik}^{-ij}\}_{k=1}^K \sim Dir(\alpha\beta_1 + N_{i1}^{-ij}, \cdots, \alpha\beta_K + N_{iK}^{-ij}, \alpha\beta_{K+1})$. Therefore, having the knowledge of $F(\pi_i|\alpha, \beta, \{N_{ik}^{-ij}\}_{k=1}^K)$, given u_{ij}, our calculation of $\Pr(s_{ij}=k)$ is equal to computing the probability of u_{ij} falling in π_i's k^{th} interval, i.e. $\Pr(\sum_{d=1}^{k-1}\pi_{id} \le u_{ij} < \sum_{d=1}^{k}\pi_{id})$ (similar case with v_{ij} to π_{jl}). This can be obtained from the fact that the set $\{u_{ij} \in [0,1]|\sum_{d=1}^{k-1}\pi_{id} \le u_{ij}\}$ can be decomposed into two disjoint sets:

$$\begin{aligned}
&\{u_{ij} \in [0,1]|\sum_{d=1}^{k-1}\pi_{id} \le u_{ij}\} \\
=&\{u_{ij} \in [0,1]|\sum_{d=1}^{k-1}\pi_{id} \le u_{ij} < \sum_{d=1}^{k}\pi_{id}\} \\
&\cup \{u_{ij} \in [0,1]|\sum_{d=1}^{k}\pi_{id} \le u_{ij}\}
\end{aligned} \tag{4.23}$$

where $\sum_{d=1}^{k}\pi_{id} \sim Beta(\sum_{d=1}^{k}\alpha\beta_d + N_{id}, \sum_{d=k+1}^{K+1}\alpha\beta_d + N_{id})$. (A similar

result was also found in page 10 of (Teh et al. 2006)). Therefore, we have:

$$\begin{aligned}
&\Pr(\sum_{d=1}^{k-1}\pi_{id} \le u_{ij} < \sum_{d=1}^{k}\pi_{id}) \\
=&\Pr(\sum_{d=1}^{k-1}\pi_{id} \le u_{ij}) - \Pr(\sum_{d=1}^{k}\pi_{id} \le u_{ij}) \\
=&I_{u_{ij}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}}(h_i^{k}, \hat{h}_i^{k})
\end{aligned} \tag{4.24}$$

Here $h_i^k = \sum_{d=1}^{k}\alpha\beta_d + N_{id}, \hat{h}_i^k = \sum_{d=k+1}^{K+1}\alpha\beta_d + N_{id}$; $I_u(a,b)$ denotes the Beta c.d.f. value with parameter a, b on u. The existence and non-negativity of $I_{u_{ij}}(u_{k-1}, \hat{u}_{k-1}) - I_{u_{ij}}(u_k, \hat{u}_k)$ is guaranteed by the fact that $\{u_{ij} \in [0,1] | \sum_{d=1}^{k}\pi_{id} \le u_{ij}\} \subseteq \{u_{ij} \in [0,1] | \sum_{d=1}^{k-1}\pi_{id} \le u_{ij}\}$ on the same π_i.

With the same notation as in Eq. (4.12), we have

$$\begin{aligned}
&\Pr(s_{ij}=k, r_{ij}=l | \\
&\quad e_{ij}, \lambda_1, \lambda_2, m_{k,l}, u_{ij}, v_{ij}, \{h_i^k\}_k, \{\hat{h}_i^k\}_k, \{h_j^k\}_k, \{\hat{h}_j^k\}_k) \\
&\propto (I_{u_{ij}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}}(h_i^k, \hat{h}_i^k)) \\
&\quad \cdot (I_{v_{ij}}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}}(h_j^l, \hat{h}_j^l)) \\
&\quad \cdot \begin{cases} m_{k,l}^{1,-e_{ij}} + \lambda_1, & e_{ij}=1; \\ m_{k,l}^{0,-e_{ij}} + \lambda_2, & e_{ij}=0. \end{cases}
\end{aligned} \tag{4.25}$$

Sampling m and β

In the Marginal conditional on u, v only method, since we have integrated out π, therefore, we follow the method similar to that of (Teh et al. 2006), and use the auxiliary variable $\boldsymbol{m}$ which is distributed as (Antoniak 1974)(Van Gael, Saatci, Teh & Ghahramani 2008):

$$p(m_{ik} = m | N_{ik}, \alpha, \boldsymbol{\beta}_k) \propto S(N_{ik}, m)(\alpha\boldsymbol{\beta}_k)^m \tag{4.26}$$

Here $S(\cdot,\cdot)$ is the Stirling number of first kind. And the parameter $\boldsymbol{\beta}$ has the posterior distribution:

$$(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K, \boldsymbol{\beta}_{K+1}) \sim Dir(m_{\cdot 1}, \cdots, m_{\cdot K}, \boldsymbol{\gamma}) \tag{4.27}$$

Here $\boldsymbol{\beta}_{K+1}$ denotes the proportion of undetected communities.

Sampling Hyper-Parameters

Hyper-parameters α and γ are the same with the other methods in this chapter. Similar as Marginal conditional on π only, in Marginal conditional on u, v only, we also use the M-H sampling to update Copula function parameter θ's value. With the proposal distribution being the prior $\mathcal{G}^*(1,1)$ (definition in Eq. (14)), we obtain the following acceptance rate:

$$A(\theta_d^*, \theta_d^{(\tau)}) = \min\left(1, \frac{\prod_{i,j} c(u_{ij}, v_{ij}|\theta^*)}{\prod_{i,j} c(u_{ij}, v_{ij}|\theta^{(\tau)})}\right) \tag{4.28}$$

4.4.3 Relations with Classical MMSB

A bivariate independence Copula function, i.e. $C(u,v) = uv$, is a uniform distribution on the region of $[0,1]\times[0,1]$. Under the case of "marginal conditional on π only", Eq. (4.14) then becomes that of $p_{ij}^{kl}(\pi_i, \pi_j) = \Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j) = \int_{\hat{\pi}_i^{k-1}}^{\hat{\pi}_i^k} \int_{\hat{\pi}_j^{l-1}}^{\hat{\pi}_j^l} \cdot 1 du dv = \pi_{ik} \cdot \pi_{jl}$. Under the case of "marginal conditional on u, v only", as $\{u_{ij}, v_{ij}\}$ are independently uniform distributed, the equation $\int_{u_{ij}} \Pr(\sum_{d=1}^{k-1} \pi_{id} < u_{ij} \leq \sum_{d=1}^{k} \pi_{id}) du_{ij} = \sum_{d=1}^{k} \pi_{id} - \sum_{d=1}^{k-1} \pi_{id} = \pi_{ik}$. (A similar result also holds for v_{ij}.) All these results are identical to that of the classical MMSB. In a sense, our model can be viewed as a generalization of MMSB.

In addition, for most Copula functions, a certain choice of parameters will result in the function equalling or approaching that of the independence Copula. As an example, when Gumbel (Nelsen 2006) Copula is used, it has its c.d.f. defined as:

$$C(u,v) = \exp\left[-\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{\frac{1}{\theta}}\right] \tag{4.29}$$

where $\theta \in [1, \infty)$. For $\theta = 1$, it becomes that of the independence Copula. Our experiments show that when the data are generated using independence Copula (i.e., classical MMSB), the recovered Gumbel Copula's parameter has a high probability of around 1.

4.4.4 Relations with Dependent Dirichlet Process

As we should note that, since the proposal of dependent Dirichlet process (DDP) (MacEachern 1999), a variety of DDP models were developed, including a recent Poission Process perspective (Lin et al. 2010) and its variants (Lin & Fisher 2012)(Foti, Futoma, Rockmore & Williamson 2013)(Chen et al. 2013).

From the dependency modeling perspective, our Copula-incorporated work achieves a similar goal to that of DDP. However, the DDP-type works concentrate on the intrinsic relations between multiple Dirichlet Processes. In our work, however, we assume Dirichlet Processes themselves are independent. The dependency is introduced at the (discrete) realizations of the multiple DPs, which are the membership indicators. Therefore, making it feasible to use Copula to model the dependency between each pair of membership indicators. This obviously cannot be achieved at the DP level, as one's relations with every other nodes share the same DP.

4.4.5 Computational Complexity Analysis

We estimate the computational complexity for each graphical model and present the result in Table 4.1. Compared to the classical models (especially the MMSB), our cMMSB$^{\pi}$ involves an additional $\mathcal{O}(Kn)$ term which refers to the sampling of the mixed membership distributions. Note that the computational time varies for different Copulas. cMMSB^{uv} requires an extra $\mathcal{O}(n^2)$ term for the u, v's sampling for each membership indicator. Each operation requires a beta c.d.f. in a tractable form.

4.5 Experiments

Here, our cMMSB's performance is compared with the classical mixed-membership stochastic blockmodel (MMSB)-type methods, including the original MMSB (Airoldi et al. 2008) and the infinite mixed-membership model (iMMM) (Koutsourelakis & Eliassi-Rad 2008). Additionally, we also com-

Table 4.1: Computational Complexity for Different Models

Models	Computational Complexity
IRM	$\mathcal{O}(K^2L)$[1]((Palla et al. 2012))
LFRM	$\mathcal{O}(K^2n^2)$ ((Palla et al. 2012))
MMSB	$\mathcal{O}(Kn^2)$ ((Kim et al. 2012))
cMMSB$^{\pi}$	$\mathcal{O}(Kn^2 + Kn) = \mathcal{O}(Kn^2)$
cMMSB^{uv}	$\mathcal{O}(Kn^2 + n^2) = \mathcal{O}(Kn^2)$

[1] $L = \sum_{i,j} \boldsymbol{e}_{ij}$, denotes the number of positive link data.

pare it with other non-MMSB approaches including the infinite relational model (IRM) (Kemp et al. 2006), the latent feature relational model (LFRM) (Miller et al. 2009) and the nonparametric metadata dependent relational model (NMDR) (Kim et al. 2012).

We independently implement the above benchmark algorithms to the best of our understanding. In order to provide a common ground for all comparisons, we make the following small variations to these algorithms: (1) In iMMM, instead of having an individual α_i value for each π_i as used in the original work, we use a common α value for all the mixed-membership distributions $\{\pi_i\}_{i=1}^n$; (2) In LFRM (Miller et al. 2009)'s implementation, we do not incorporate the metadata information into the interaction data's generation, but use only the binary interaction information.

As the predict ability is one important property of the model, we use a ten-fold cross-validation to complete this task, where we randomly select one out of ten for each node's link data as test data and the others as training data. Each model is run for 50 times for fair comparison and the corresponding statistics (including the mean and standard deviation) are reported.

4.5.1 Synthetic Data

We first perform the synthetic data exploration as a pilot study. In addition to the ones associated with the Copula function, the rest of the variables are

Table 4.2: Model Performance (Mean $\mp$ Standard Deviation) on Synthetic Data of Full Correlation.

	Train error	Test error	Test log likelihood	AUC
IRM	0.104 ∓ 0.010	0.105 ∓ 0.013	-90.060 ∓ 8.217	0.874 ∓ 0.026
LFRM	0.094 ∓ 0.003	0.113 ∓ 0.014	-99.997 ∓ 10.8921	0.872 ∓ 0.023
MMSB	$\mathbf{0.024 \mp 0.000}$	0.125 ∓ 0.000	-104.107 ∓ 0.265	0.851 ∓ 0.001
iMMM	0.027 ∓ 0.0002	0.121 ∓ 0.0003	-101.497 ∓ 0.203	0.862 ∓ 0.001
cMMSB$^{\pi}$	0.033 ∓ 0.001	$\mathbf{0.088 \mp 0.000}$	$\mathbf{-82.625 \mp 0.128}$	$\mathbf{0.890 \mp 0.000}$
cMMSB^{uv}	0.042 ∓ 0.000	0.093 ∓ 0.000	-85.951 ∓ 0.084	0.889 ∓ 0.001
cMMSB$^{\pi}$(P)[1]	0.034 ∓ 0.001	0.089 ∓ 0.000	-83.264 ∓ 0.105	0.894 ∓ 0.001
cMMSB^{uv}(P)[1]	0.0497 ∓ 0.001	0.091 ∓ 0.001	-83.124 ∓ 0.046	0.895 ∓ 0.007

[1] This is under the situation of Partial Correlation, i.e., we are using two Copula functions in different subgroups.

generated in accordance with (Airoldi et al. 2008)(Newman & Girvan 2004). We use $n = 50$, and hence E is a 50×50 asymmetric and binary matrix. The parameters are set up such that 50 nodes are partitioned into 4 subgroups, with each subgroup having $20, 13, 9, 8$ number of nodes, respectively. The mixed-membership distribution of each group and the whole role-compatibility matrix are displayed in Fig. 4.2 and Fig. 4.3, respectively. Thus, the generated synthetic data forms as one block diagonal matrix, with the outliers existed.

$$\begin{vmatrix} 0.9 & 0.1 & 0 & 0 \\ 0 & 0.9 & 0.1 & 0 \\ 0.1 & 0.05 & 0.85 & 0 \\ 0.1 & 0.05 & 0.05 & 0.8 \end{vmatrix}$$

Figure 4.2: Mixed-membership Distribution

$$\begin{vmatrix} 0.95 & 0.05 & 0 & 0 \\ 0.05 & 0.95 & 0.05 & 0 \\ 0.05 & 0 & 0.95 & 0 \\ 0 & 0.05 & 0 & 0.95 \end{vmatrix}$$

Figure 4.3: Role-compatibility Matrix

Full Correlation - Single Copula on All Nodes in Link Prediction

We incorporate a single Gumbel Copula (with parameter $\theta = 3.5$) on every interaction to generate all membership indicator pairs. The corresponding average values are shown in Table 4.2. The definitions for the comparison indicator such as train error, test error, test log likelihood and AUC can be referred to (Kim et al. 2012).

An interesting part of our results is that we find the IRM is slightly better than the LFRM and MMSB, we explain this to the "blockness" of the synthetic data. In terms of train error, our model is comparative to

other MMSB-type models, which in general outperforms IRM and LFRM. On the predictability measures on test error, test log likelihood and AUC, both our cMMSB$^{\pi}$ and cMMSB^{uv} outperform all other MMSB and non-MMSB benchmarks.

Another interesting comparison is the posterior predictive distribution on the train data, and we have shown its results and detailed discussion in Fig. 4.4. The corresponding value is calculated as the average value of the second half of the samples in one chain, as the first half is set being the "burn in" stage. The darker of the pointer stands for the larger value close to 1, and vice versa.

The original data is a block diagonal matrix, with some outliers existed as the black points. For the IRM model, its result is composed of rectangular zones. One value is presented in each rectangular. This simplified and "smoothed" version is due to the single membership representation for one node, it cannot distinguish the random distributed points. Comparing to this, the LFRM provides a larger amount of values to select from. This enables the model to place different values on one rectangular zone, especially each node is meant to be line-shaped colors, which is in consistent with the one latent feature vector for one node representation. However, it still fails to detect the random points. MMSB and iMMM successfully capture the random points. What is more, we find our cMMSB$^{\pi}$ and cMMSB^{uv} models partition the relational data the best.

Partial Correlation - Multiple Copulas in Subgroup Structure

We also have an additional test case and integrate two Gumbel Copula functions in the modelling. The first 20 nodes form a correlated subgroup and share one Copula function, while the other Copula function is applied on the rest of the interactions. The performance on this partial correlation data is shown in the bottom two rows of Table 4.2.

While using this model on a partial correlation dataset, we obtain the 95% Confidence Interval for both of the recovered θ_1 and θ_2 displayed in

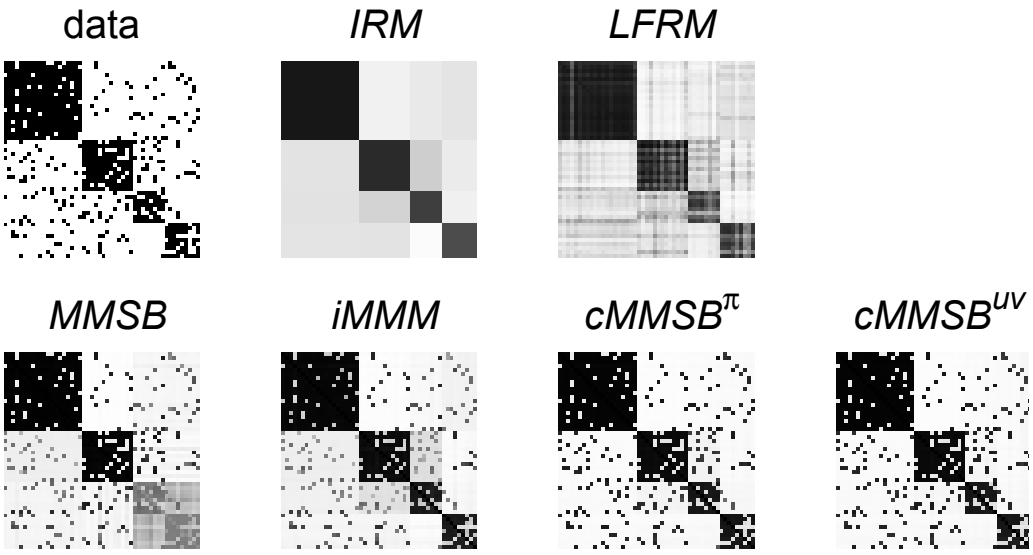

Figure 4.4: Comparison of Models' Posterior Predictive Distribution on the Training data.

Table 4.3. We can see that our model can distinguish between the correlated and independent cases, where the recovered value of θ_2 is much closer to 1. In Figure 4.4, the IRM tends to "re-construct" the training relational data in single-density rectangles, while the re-construction in the LFRM contains strip-type variants. All of the bottom 4 plots are MMSB-type model, while they tend to precisely re-construct the relational data (including the noise data). Also, we have found our models achieve better performance against MMSB and iMMM.

Table 4.3: θ's 95% Confidence Interval

Models	s-cMMSB	s-ciMMM	Ground-truth
θ_1	4.19 ∓ 0.91	3.23 ∓ 1.22	3.5
θ_2	1.42 ∓ 0.23	2.39 ∓ 0.48	1.0

4.5.2 Real-world Datasets for Link Prediction

We analyse three real-world datasets: the NIPS Co-authorship dataset, the MIT Reality Mining dataset (Eagle & Sandy 2006) and the Lazega-lawfirm dataset (Lazega 2001). Under the same execution setting as in synthetic data, we show the detailed results in Table 4.6.

NIPS Co-authorship Dataset

We use the co-authorship as a relation from the proceeding of the Neural Information Processing Systems (NIPS) conference for the years 2000-2012. Due to the sparse nature of the co-authorships, we observe the authors' activities in all the 13 years (i.e. regardless of the time factor) and set the relational data to 1 if the two corresponding authors have co-authored for no less than 2 papers, which remove some of the "by chance" co-authorships. Further, the author with less than 4 relationships with others are considered "inactive" and hence have been manually removed. Thus, a 92×92 symmetric and binary matrix is obtained.

On this dataset, no pre-defined group information is obtained in advance. Thus, we consider it as full-correlation case and use one Gumbel Copula function to model all the interactions.

MIT Reality Dataset

From the MIT Reality Mining (Eagle & Sandy 2006), we use the subjects' proximity dataset, where weighted links indicate the average proximity from one subject to another at work. We then "binarize" the data, in which we set the proximity value larger than 10 minutes per day as 1, and 0 otherwise. Therefore, a 94×94 asymmetric and binary matrix is obtained.

The dataset are roughly divided into four groups: Sloan Business School students (Sloan), lab faculty, senior students with more than 1 year in the lab and junior students. In our experiment, we only apply the Gumbel Copula function to the Sloan portion of the students to encourage similar mixture

Table 4.4: Model Performance on NIPS Co-author dataset(Mean $\mp$ Standard Deviation)

Models	Train error	Test error	Test log likelihood	AUC
IRM	0.032 ∓ 0.000	0.042 ∓ 0.001	-135.047 ∓ 7.382	0.890 ∓ 0.016
LFRM	0.048 ∓ 0.080	0.024 ∓ 0.074	-105.217 ∓ 179.551	0.935 ∓ 0.167
MMSB	0.013 ∓ 0.004	0.030 ∓ 0.006	-86.213 ∓ 10.126	0.952 ∓ 0.022
iMMM	$\mathbf{0.006 \mp 0.002}$	0.025 ∓ 0.004	-83.426 ∓ 9.429	0.957 ∓ 0.016
cMMSB$^{\pi}$	0.007 ∓ 0.004	$\mathbf{0.023 \mp 0.004}$	-83.426 ∓ 9.428	0.957 ∓ 0.016
cMMSB^{uv}	0.010 ∓ 0.005	0.024 ∓ 0.007	$\mathbf{-83.426 \mp 9.429}$	$\mathbf{0.958 \mp 0.015}$

Table 4.5: Model Performance on MIT Reality dataset(Mean $\mp$ Standard Deviation)

Models	Train error	Test error	Test log likelihood	AUC
IRM	0.063 ∓ 0.000	0.067 ∓ 0.000	-133.804 ∓ 1.127	0.826 ∓ 0.005
LFRM	0.040 ∓ 0.002	0.063 ∓ 0.004	-143.607 ∓ 10.059	0.853 ∓ 0.018
MMSB	0.026 ∓ 0.011	0.072 ∓ 0.004	-129.436 ∓ 7.655	0.856 ∓ 0.018
iMMM	0.030 ∓ 0.006	0.063 ∓ 0.002	-126.788 ∓ 3.477	0.862 ∓ 0.012
NMDR	0.039 ∓ 0.004	0.067 ∓ 0.001	-139.523 ∓ 2.937	0.857 ∓ 0.014
cMMSB$^{\pi}$	$\mathbf{0.025 \mp 0.002}$	0.049 ∓ 0.002	-125.388 ∓ 3.269	$\mathbf{0.879 \mp 0.016}$
cMMSB^{uv}	0.028 ∓ 0.004	$\mathbf{0.044 \mp 0.002}$	$\mathbf{-123.388 \mp 3.125}$	0.874 ∓ 0.036

Table 4.6: Model Performance on Lazega dataset(Mean $\mp$ Standard Deviation)

Models	Train error	Test error	Test log likelihood	AUC
IRM	0.099 ∓ 0.000	0.105 ∓ 0.001	-201.791 ∓ 3.350	0.706 ∓ 0.017
LFRM	0.057 ∓ 0.002	0.105 ∓ 0.006	-222.592 ∓ 16.199	0.817 ∓ 0.020
MMSB	0.039 ∓ 0.007	0.091 ∓ 0.003	-212.126 ∓ 3.215	0.799 ∓ 0.010
iMMM	0.049 ∓ 0.007	0.110 ∓ 0.003	-202.715 ∓ 5.308	0.807 ∓ 0.014
NMDR	0.064 ∓ 0.006	0.113 ∓ 0.002	-207.719 ∓ 3.475	0.829 ∓ 0.011
cMMSB$^{\pi}$	$\mathbf{0.025 \mp 0.005}$	$\mathbf{0.102 \mp 0.006}$	$\mathbf{-201.015 \mp 5.217}$	$\mathbf{0.827 \mp 0.015}$
cMMSB^{uv}	0.028 ∓ 0.004	0.114 ∓ 0.002	-204.029 ∓ 9.546	0.822 ∓ 0.017

membership indicators.

Lazega Law Dataset

The lazega-lawfirm dataset (Lazega 2001) is obtained from a social network study of corporate located in the north-eastern part of U.S. in 1988 - 1991. The dataset contains three different types of relations: co-work network, basic advice network and friendship network, among the 71 attorneys, of which the element are labeled as 1 (exist) or 0 (absent).

Since no group information is obtained in this dataset, we use the same setting as in NIPS co-authorship dataset as one Gumbel Copula function is used for all the interactions.

General Performance

From these reported statistics as shown in Table 4.6, we can see that our methods (cMMSB$^{\pi}$, cMMSB^{uv}) obtain the best performance in these 3 datasets, amongst all other models. Although iMMM can achieve smallest train error in the NIPS co-author dataset, our cMMSB's predictability is better than iMMM and the others. On the MIT reality and Lazega-lawfirm datasets, our cMMSB can achieve at least 1% improvement on the AUC score. On the performance comparison of our two different sampling schemes cMMSB$^{\pi}$ and cMMSB^{uv}, we find they achieve similar results, which is within our expectation.

Our cMMSB$^{\pi}$, cMMSB^{uv} beat both MMSB-liked models and non-MMSB models since a hidden intra-group correlation has been adaptively utilized here. As its widely existence in social network, this additional information is expected to contribute to the model's performance, which is verified in our experiment.

4.6 Summary

In this chapter, we have tried to integrate the Copula function to fully describe the coupling relations within the communities of the networks, which is to greatly complement the mixed-membership stochastic blockmodel which essentially treats nodes independently. The Copula function is used to represent the correlation between the pair of membership indicators, while keeping the membership indicators' marginal distribution invariant. The results show that, using both synthetic and real data, our Copula-incorporated MMSB, i.e., cMMSB, is effective in learning the community structure and predicting the missing links.

4.7 Limitation & Future Work

As this chapter discusses an elegant integration of the Copula function into the Dirichlet Process, the main limitation may be the complicated generative process and its corresponding inference schedule. Although we have tried two different inference methods to address the problem, the computational cost is still larger than the classical ones. Also, as the single Copula function may be insufficient to capture the complex coupling relations within the communities, multiple-Copula functions may be used as a trial.

Besides this copula integration method, there are other ways in considering the subgroup correlation, such as conditioning on the value of g_{ij} or simply using the logistic function to model the observation. These interesting methods are intuitive and easy implemented, however, can not provide find-grained control of the effect that g_{ij} has on the e_{ij}. Also, we should note that although s_{ij} and r_{ij} are generated independently, their corresponding product have been "re-measured" in a "copula" way (see the second inference method). From this perspective, we are arguing that these membership indicator pair preserve the "coupling relation" we are modelling.

The focus here is on using one Copula function (Gumbel Copula) to explore the within communities' coupling relations. A natural extension is

to use multiple Copulas on different subgroups, as the various types of Copula functions provide multiple options to capture various dependencies. Inspired by the multiple kernel learning, a linear combination of Copula functions used in cMMSB is also a promising direction for future research.

Apart from these modelling choices, there is also interests in more applications involving the Copula function. For instance, two Indian Buffet Processes can also be described by using the Copula function, as well as the Hierarchical Dirichlet Process - Hidden Markov Model. This modelling within this MMSB framework can be regarded as one pilot work to provide a different way in introducing the Copula function to the graphical model.

Chapter 5

Learning Relational Models by Efficiently Involving Node Information in a Network

5.1 Introduction

Community detection and network partitioning is an emergent topic in various areas including social-media recommendation (Tang & Liu 2010), customer partitioning, social network analysis, and partitioning protein interaction network tasks (Girvan & Newman 2002) (Fortunato 2010). Many models have been proposed in recent years to address this problem by using link data (e.g. a person's view towards others). Some examples include the stochastic blockmodel (Nowicki & Snijders 2001) and in the case of infinite communities, the infinite relational model (IRM) (Kemp et al. 2006), both aiming at partitioning a network of nodes into different groups based on their pairwise, directional binary observations. In most existing approaches, the "inter-nodes" link data is a lone contributor towards the understanding of the insights of social structures.

On the other hand, the "intra-nodes" information is a vital source of additional information to complement the link information. Let us take the

Lazega dataset (Lazega 2001) (detailed in the experimental section of this chapter), which is a social network within a lawyer firm, as an example. The node (i.e. attorney) here contains information such as ages, offices (Boston, Hartford or Providence), and law schools (Harvard, Yale, Ucon or other). Naturally, the attorneys with similar information (e.g. the same office) tend to have relationships, and/or belong to same community. This kind of **dependent coupling** is no doubt to facilitate us with a much more complete understanding of the network.

While some recent efforts have been directed to incorporate the node information, they all face several shortcomings mainly in terms of appropriateness and efficiency. For example, in terms of appropriateness, in LFRM (Miller et al. 2009), although the direct and linear combination of node information and the latent feature have experimentally demonstrated its effectiveness in link prediction, it is hard to interpret the recovered features and their related social structure (also stated in (Kim et al. 2012)). In terms of efficiency, taking NMDR (Kim et al. 2012) as example, the logistic-normal transform was employed to integrate the node information into each node's mixed-membership distribution. However, this integration complicates the original structure and results in non-conjugacy during the inference.

Two major branches of relational models have been developed in the last few years, namely the MMSB (Airoldi et al. 2008) and LFRM (Miller et al. 2009), where community memberships are modelled as mixed memberships and latent features respectively. In order to demonstrate the generality of our method, we have individually adapted our method to both of these frameworks, and have produced two distinct models, which is the central theme of this chapter: the node-information involved mixed-membership model (niMM) and the node-information involved latent-feature model (niLF). In both cases, methods similar to the stick-breaking process (Sethuraman 1994)(Teh et al. 2007) are proposed to model the unknown number of communities. In particular, niMM successfully obtains the conjugate property during the MCMC inference procedure. As discussed later, through these efforts, the

existing models (MMSB and LFRM) can be seen as special cases of our proposed models. In this way, our models capture much richer information embedded in a network; hence, they result in better performance in modeling the communities' memberships as illustrated in the experiments.

In summary, our contributions can be stated as: 1) we have naturally extended the existing benchmark models (i.e. MMSB and LFRM) to incorporate the nodes' information. The experimental results seem quite promising while the nodes' information is closely related to the link data; 2) our extension to MMSB has retrieved the conjugate property during the MCMC inference, which mixes much faster in the Markov Chain than the previous approaches. Also, we find that in the experiments, our method converges much earlier than the previous one; 3) our model is under the Bayesian Nonparametrics setting (achieved through the methods similar to the stick-breaking constructions), which can deal with the problem of an unknown number of communities.

The rest of the chapter is organized as follows. We describe both our niMM and niLF models in details, as well as the detail inference procedure and a "collapsed" inference discussion of niMM. We also include the model's computational complexity analysis in the same section. In the experimental section, we compare our methods with the previous work to validate the models performances. The conclusions and future work are given in the last section.

5.2 Literature Review of Stick-breaking Process

The stick-breaking method (Sethuraman 1994)(Ishwaran & James 2001) has provided an explicit construction of a draw G from a Dirichlet process:

$$\begin{aligned} G &= \sum_{k}^{\infty} \pi_k \delta_{\theta_k}, \pi_k = \psi_k \prod_{l=1}^{k-1}(1-\psi_l), \\ \psi_k &\overset{iid}{\sim} \text{Beta}(1,\gamma), \theta_k \overset{iid}{\sim} G_0. \end{aligned} \tag{5.1}$$

The concentration parameter γ controls the diversity of θ in G, whereas G_0 is regarded as the base measure generating $\{\theta_k\}_{k=1}^{\infty}$. A larger γ encourages the weights distribution to be more "flat", whereas a smaller γ stimulates the weights to be "sharper", i.e., only a few weights have appreciable values and the others are relatively small. As an indication of the importance of this concentration parameter γ, a vague gamma prior distribution is usually placed on it.

Based on this ingenious construction, more flexible constructions have been proposed, the recent examples being the logistic stick-breaking process (Ren et al. 2011), the probit stick-breaking process (Rodriguez & Dunson 2011), the kernel stick-breaking process (Dunson & Park 2008), and the discrete infinite logistic normal process (Paisley, Wang & Blei 2012). While being elastic in describing the Bayesian Nonparametric prior in different situations, one common problem is that they cannot form a prior-posterior conjugate design, which caused difficulties for both the MCMC sampling inference (using Metropolis-Hastings Sampling instead would greatly slow down the mixing rate) and variational inference (having to find an approximate distribution to replace this distribution).

5.3 Generative Model

Figure 5.1 depicts the graphical models of all the variables used in our work. Observational variables are colored in grey. $\{\phi_i\}_{i=1}^{n}$ is the nodes' attributes

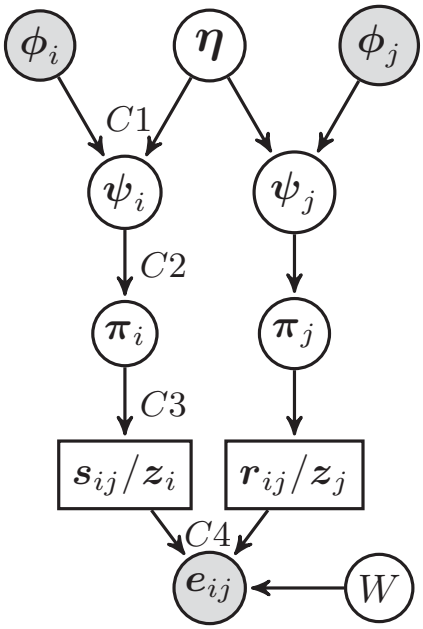

Figure 5.1: The generative model for the niMM and niLF models.

information (transformed into one binary vector), and $\{e_{ij}\}_{i,j}$ stands for the observational link data. $C1 - C4$ represent four conditional distributions in two different forms as shown in the niMM and niLF sections, respectively. s_{ij} and r_{ij} in the rectangular nodes represent the latent label in niMM, and z_i and z_j are in the niLF context. As previously discussed, node information is incorporated into both branches of the relational models: iMMM and LFRM. Therefore, we illustrate both in the same figure, as most nodes are common to both graphical models.

5.3.1 Node-information Involved Mixed-Membership Model

The generative process for the node-information involved mixed membership (niMM) model is defined as follows (W.l.o.g. $\forall i, j = 1, \ldots, n, k \in N^+$):

$C1$, $\psi_{ik} \sim \text{Beta}(1, \prod_f \eta_{fk}^{\phi_{if}})$;

$C2$, $\pi_{ik} = \psi_{ik} \prod_{l=1}^{k-1}(1 - \psi_{il})$;

$C3$, $s_{ij} \sim \text{Multi}(\pi_i), r_{ij} \sim \text{Multi}(\pi_j)$;

$C4$, $e_{ij} \sim \text{Bernoulli}(W_{s_{ij}r_{ij}})$.

Here, $C1$ and $C2$ constitute the stick-breaking representation for our mixed-membership distribution π_i, which is similar to that of the Dirichlet process. While the Dirichlet process employs one single γ parameter to finish its stick-breaking construction, our representation uses different values for each component. The values are computed through exponential form $\eta_{fk}^{\phi_{if}}$ to further facilitate the conjugate design. $C3$ and $C4$ correspond to the membership indicator and link data generation, which follows the procedure as in Section 2.3.1.

On $C1$, we replace the fixed γ parameter in the stick-breaking process with $\prod_f \eta_{fk}^{\phi_{if}}$, where the positive, importance indicator η_{fk} is given a vague gamma prior $\eta_{fk} \sim \text{Gamma}(\alpha_\eta, \beta_\eta)$. Our method can successfully integrate the node information into the node's mixed-membership distribution and enjoy the conjugate property during the inference procedure. On the other hand, the previous approach (Kim et al. 2012)(Kim & Sudderth 2011) uses the logistic normal distribution (with the mean value being the linear sum (i.e., $\sum_f \phi_{if}\eta_{fk}$)) to construct a stick-breaking weight ψ_{ik}, which makes the inference inefficient (i.e. slow mixing rate during the MCMC sampling).

We again use the attribute *age* (which will be "binarized" before use) in the Lazega dataset to further explain the importance indicator η_{fk} used in $C1$. W.l.o.g., we let f_0^{th} column of ϕ matrix denotes the age attribute, $\phi_{if_0} = 1$ implies that node i has $age > 40$ (in our experimental setting), and 0 otherwise. From Equation $C1$, one can easily see that when $\eta_{f_0k} \ll 1$, *age* would largely increase its impact on the k^{th} community. Likewise, $\eta_{f_0k} \gg 1$ reduces the influence of the *age* attribute on the k^{th} community. $\eta_{f_0k} = 1$ means that *age* does not have an impact on the k^{th} community at all. Also, $\phi_{if_0} = 0$ makes *age* of the node i neutral towards all other communities.

Both the importance indicator η_{fk} and stick-breaking weight ψ_{ik} can enjoy the conjugate property. More specifically, the distributions of η_{fk}, ψ_{ik} are:

$$\begin{aligned} p(\eta_{fk}|\alpha_\eta, \beta_\eta) &\propto \eta_{fk}^{\alpha_\eta - 1} e^{-\beta_\eta \eta_{fk}}; \\ p(\psi_{ik}|\eta_{\cdot k}, \phi_i) &\propto \left[\prod_f \eta_{fk}^{\phi_{if}}\right] \cdot (1-\psi_{ik})^{\prod_f \eta_{fk}^{\phi_{if}} - 1}. \end{aligned} \tag{5.2}$$

Thus, the posterior distribution of η_{fk} becomes:

$$\begin{aligned} & p(\eta_{fk}|\alpha_\eta, \beta_\eta, \psi_{\cdot k}, \phi) \propto p(\eta_{fk}|\alpha_\eta, \beta_\eta) \prod_i p(\psi_{ik}|\phi_i, \eta_{\cdot k}) \\ \propto & \eta_{fk}^{\alpha_\eta + \sum_i \phi_{if} - 1} e^{-(\beta_\eta - \sum_i \phi_{if} \ln(1-\psi_{ik}) \prod_{F \neq f} \eta_{Fk}^{\phi_{iF}})\eta_{fk}} \\ \Longrightarrow & \eta_{fk} \sim \text{Gamma}(\alpha_\eta + \sum_i \phi_{if}, \beta_\eta - \sum_i \phi_{if} \ln(1-\psi_{ik}) \prod_{q \neq f} \eta_{qk}^{\phi_{iq}}) \end{aligned} \tag{5.3}$$

The joint probability of $\{s_{ij}, r_{ji}\}_{j=1}^n$ becomes:

$$p(\{s_{ij}\}_{j=1}^n, \{r_{ji}\}_{j=1}^n | \psi_{i\cdot}) \propto \prod_{k=1}^K \left[\psi_{ik}^{N_{ik}} (1-\psi_{ik})^{\sum_{l=k+1}^K N_{il}}\right] \tag{5.4}$$

here $N_{ik} = \#\{j : s_{ij} = k\} + \#\{j : r_{ji} = k\}$.

The posterior distribution of ψ_{ik} becomes:

$$\begin{aligned} & p(\psi_{ik}|\phi, \eta_{\cdot k}, \{s_{ij}, r_{ji}\}_{j=1}^n) \\ \propto & \psi_{ik}^{N_{ik}} (1-\psi_{ik})^{\sum_{l=k+1}^K N_{il} + \prod_f \eta_{fk}^{\phi_{if}} - 1} \\ \Longrightarrow & \psi_{ik} \sim \text{Beta}(N_{ik} + 1, \sum_{l=k+1}^K N_{il} + \prod_f \eta_{fk}^{\phi_{if}}) \end{aligned} \tag{5.5}$$

The posterior distribution of ψ_{ik} in Eq. (5.5) is consistent with the result in (Ishwaran & James 2001)(Kalli et al. 2011), where their result is conditioned on a single concentration parameter α instead of $\prod_f \eta_{fk}^{\phi_{if}}$.

Another interesting comparison is the placing of prior information for communities within different models. In iMMM, although the author claimed to use different α_i to model individual π_i, the stick-breaking weights $\{\psi_{ik}\}_{k=1}^\infty$ within one π_i are generated identically, i.e., from beta$(1, \alpha_i)$. This is obviously insufficient as each community may expect an individual prior in real

application. Accordingly, NMDR has incorporated node information using a logistic normal function, as stated above. In a way, this approach has further generalised the model, such that each ψ_{ik} differs in their distributions.

Despite the model relaxation, empirical results show that NMDR has a slow convergence. It is therefore imperative for us to search for a more efficient way to incorporate the node information. Compared to iMMM, our niMM model replaces the simple set $\{\alpha_i\}$ with $\prod_f \eta_{fk}^{\phi_{if}}$ for the generation of ψ_{ik}. Its conjugate property makes our model appealing in terms of mixing efficiency, which is confirmed in the results shown in the experimental section. What is more, our model can be seen as a natural extension of the popular iMMM model. By letting $\eta_{fk} = \alpha^{1/F}$ and $\phi_{if} = 1$, we obtain the classical iMMM. This makes sense, as without the presence of metadata, each feature is assumed to be counted equally, which implies that the model becomes the classical iMMM.

5.3.2 Node-information Involved Latent Feature Model

The generative process for the node-information involved latent feature (niLF) model is defined as follows:

$C1$, $\psi_{ik} \sim \text{Beta}(\prod_f \eta_{fk}^{\phi_{if}}, 1)$;

$C2$, $\pi_{ik} = \prod_{l=1}^{k} \psi_{il}$;

$C3$, $z_{ik} \sim \text{Bernoulli}(\pi_{ik})$;

$C4$, $e_{ij} \sim \text{Bernoulli}(\frac{1}{1+\exp(-z_i W z_j^T)})$.

$C1$ and $C2$ here also constitute our specialized stick-breaking representation π_i. However, we should note that these two are different from those of the niMM model while here they are based on the traditional stick-breaking process for the Indian Buffet Process (Sethuraman 1994)(Teh et al. 2007). The π_is are used to generate the latent feature matrix z in $C3$. $C4$ corresponds

to the link data generation, which is the same as the LFRM model. Similar to the niMM model, our work can be seen as an extension of the traditional LFRM (Miller et al. 2009).

However, the structure of the stick-breaking representation in our niLF model differs from that of the LFRM model. In our niLF model, each i^{th} node's latent feature is motivated by their own stick-breaking representation π_i, i.e., there are $\boldsymbol{n}$ stick-breaking representations in total. In this way, the individual node information of node i is contained in each corresponding representation π_i, which will consequently be reflected in the latent feature. On the contrary, the LFRM model uses one specialized beta process π as the underlying representation for all the n nodes' latent feature z. This process can be easily marginalized out π_i, beneficiated from the Beta-Bernoulli conjugacy (Thibaux & Jordan 2007).

We use the new transform, i.e., $\prod_f \eta_{fk}^{\phi_{if}}$, as the mass parameter (Thibaux & Jordan 2007) in the construction of the stick-breaking representation, as stated in $C1$. The importance indicator η here plays an opposite role when compared to the niMM model, i.e., a larger value of η_{fk} would make the presence of attribute f promote the k^{th} community.

An interesting notation is that the stick-breaking representations in both our niMM and niLF models are no longer the Dirichlet process and Beta process individually, as the single valued α parameter is replaced by a set of individually-different valued $\{\prod_f \eta_{fk}^{\phi_{if}}\}$.

5.4 Inference

5.4.1 Informative Mixed Membership Model

In niMM's sampling, the variables of interest in our slice sampling are: node information weight $\{\eta_{fk}\}_{f,k}$, stick-breaking weight $\{\psi_{ik}\}_{i,k}$, latent feature indicator $\{s_{ij}, r_{ij}\}_{i,j}$, compatibility value W_{kl} and the hyper-parameters. Also, we discuss here the Beta Distribution as the generation distribution and the other ones can be trivially derived.

Sampling η_{fk}

$\forall f, k, \eta_{fk}$'s posterior distribution relies on node information $\{\phi_{if}\}_{i=1}^{n}$, stick-breaking weights $\{\psi_{ik}\}_{i=1}^{n}$, the other attribute importance indicator $\{\eta_{Fk}\}_{F\neq f}$, and its hyper-parameters α_η, β_η.

$$\eta_{fk} \sim \text{Gamma}(\alpha_\eta + \sum_i \phi_{if}, \beta_\eta - \sum_i \phi_{if} \ln(1-\psi_{ik}) \prod_{F\neq f} \eta_{Fk}^{\phi_{iF}}) \tag{5.6}$$

Sampling ψ_{ik}

$\forall i, k, \psi_{ik}$'s posterior distribution relies on $\{N_{ik}\}_{k=1}^{n}$, $\{\eta_{fk}\}_{f,k}$, $\{\phi_{if}\}_{q=1}^{q}$.

$$\psi_{ik} \sim \text{Beta}(N_{ik}+1, \sum_{l=k+1}^{K} N_{il} + \prod_f \eta_{fk}^{\phi_{if}}) \tag{5.7}$$

Sampling s_{ij}, r_{ij}

$$\Pr(e_{ij}|Z_{\setminus e_{ij}}, \alpha_W, \beta_W) = \frac{m_{kl}^{1,-e_{ij}} + \alpha_W}{m_{kl}^{-e_{ij}} + \alpha_W + \beta_W} \tag{5.8}$$

here we assume $s_{ij}=k, r_{ij}=l$, and $m_{kl}^{1,-e_{ij}} = \sum_{i'j'\neq ij, s_{i'j'}=k, r_{i'j'}=l} e_{i'j'}$, $m_{kl}^{-e_{ij}} = \sum_{i'j'\neq ij, s_{i'j'}=k, r_{i'j'}=l} 1$

Thus, we get:

$$\Pr(s_{ij}=k, r_{ij}=l) \propto \pi_{ik}\pi_{jl} \cdot \frac{m_{kl}^{1,-e_{ij}} + \alpha_W}{m_{kl}^{-e_{ij}} + \alpha_W + \beta_W} \tag{5.9}$$

When we sample $K+1$ to s_{ij} or r_{ij}, we need to re-sample the corresponding $\{\eta_{fK+1}\}_{f=1}^{F}, \psi_{iK+1}$(or ψ_{jK+1}) to the new $(K+1)^{\text{th}}$ component.

Sampling Hyper-parameters α_η, β_η, α_W, β_W

The hyper-parameters we are sampling are α_η, β_η, α_W, β_W.

For α_η, we set a vague prior Gamma$(\alpha_{\alpha_\eta}, \beta_{\alpha_\eta})$:

$$\begin{aligned} & p(\alpha_\eta|\{\eta_{fk}\}_{f,k}, \beta_\eta, \alpha_{\alpha_\eta}, \beta_{\alpha_\eta}) \\ \propto & \prod_{f,k} \left[\frac{{\beta_\eta}^{\alpha_\eta}}{\text{Gamma}(\alpha_\eta)} \eta_{fk}^{\alpha_\eta - 1}\right] \cdot {\alpha_\eta}^{\alpha_{\alpha_\eta}-1} e^{-\beta_{\alpha_\eta}\alpha_\eta} \end{aligned} \tag{5.10}$$

As Eq. (5.10) is log-concave in α_η, we use Adaptive Rejection Sampling (ARS) to finish its update.

For β_η, we set a vague prior Gamma$(\alpha_{\beta_\eta}, \beta_{\beta_\eta})$:

$$\begin{aligned} & p(\beta_\eta | \{\eta_{fk}\}_{f,k}, \alpha_\eta, \alpha_{\beta_\eta}, \beta_{\beta_\eta}) \\ \propto & \prod_{f,k} \left[{\beta_\eta}^{\alpha_\eta} e^{-\beta_\eta \eta_{fk}} \right] \cdot \beta_\eta^{\alpha_{\alpha_\eta}-1} e^{-\beta_{\alpha_\eta}\beta_\eta} \\ \propto & {\beta_\eta}^{KF\cdot\alpha_\eta+\alpha_{\alpha_\eta}-1} \cdot e^{-(\sum_{f,k}\eta_{fk}+\beta_{\alpha_\eta})\beta_\eta} \\ \Longrightarrow & \beta_\eta \sim \text{Gamma}(KF\cdot\alpha_\eta + \alpha_{\alpha_\eta}, \sum_{f,k}\eta_{fk} + \beta_{\alpha_\eta}) \end{aligned} \tag{5.11}$$

α_W and β_W is similar as above, we set a vague prior Gamma$(\alpha_{\alpha_W}, \beta_{\alpha_W})$:

$$\begin{aligned} & p(\alpha_W | \{W_{kl}\}_{k,l}, \beta_W, \alpha_{\alpha_W}, \beta_{\alpha_W}) \\ \propto & \prod_{k,l} \left[\frac{{\beta_W}^{\alpha_W}}{\text{Gamma}(\alpha_W)} W_{kl}^{\alpha_W-1} \right] \cdot \alpha_W^{\alpha_{\alpha_W}-1} e^{-\beta_{\alpha_W}\alpha_W} \end{aligned} \tag{5.12}$$

As Eq. (5.12) is log-concave in α_W, we use Adaptive Rejection Sampling (ARS) to finish its update.

For β_W, we set a vague prior Gamma$(\alpha_{\beta_W}, \beta_{\beta_W})$:

$$\begin{aligned} & p(\beta_W | \{W_{kl}\}_{k,l}, \alpha_W, \alpha_{\beta_W}, \beta_{\beta_W}) \\ \propto & \prod_{k,l} \left[{\beta_W}^{\alpha_W} e^{-\beta_W W_{kl}} \right] \cdot \beta_W^{\alpha_{\alpha_W}-1} e^{-\beta_{\alpha_W}\beta_W} \\ \propto & {\beta_W}^{K^2\cdot\alpha_W+\alpha_{\alpha_W}-1} \cdot e^{-(\sum_{kl}W_{kl}+\beta_{\alpha_W})\beta_W} \\ \Longrightarrow & \beta_W \sim \text{Gamma}(K^2\cdot\alpha_W + \alpha_{\alpha_W}, \sum_{kl}W_{kl} + \beta_{\alpha_W}) \end{aligned} \tag{5.13}$$

5.4.2 Informative Latent Feature model

In niLF's sampling, the variables of interest in our slice sampling are: node information weight $\{\eta_{fk}\}_{f,k}$, stick-breaking weight $\{\psi_{ik}\}_{i,k}$, latent feature indicator $\{s_{ij}, r_{ij}\}_{i,j}$, compatibility vale W_{kl} and the hyper-parameters.

Sampling $\boldsymbol{\eta}$

$$\eta_{fk} \sim \text{Gamma}(\alpha_\eta + \sum_i \phi_{if}, \beta_\eta - \sum_i \phi_{if} \ln \psi_{ik} \prod_{F \neq f} \eta_{Fk}^{\phi_{iF}}) \tag{5.14}$$

Sampling ψ_{ik}

We use Metropolis-Hastings Sampling to obtain the ψ_{ik}'s value, so the acceptance ratio becomes that of ($\pi_{ik} = \prod_{l=1}^{k} \psi_{il}$):

$$A(\psi_{ik}^*, \psi_{ik}^{(\tau)}) = \frac{\pi_{ik}^{*,z_{ik}}(1-\pi_{ik}^*)^{1-z_{ik}}}{\pi_{ik}^{(\tau),z_{ik}}(1-\pi_{ik}^{(\tau)})^{1-z_{ik}}} \tag{5.15}$$

Sampling u_i

We introduce an auxiliary slice variable u_i for each node i:

$$u_i | z_i, \pi \sim \text{Uniform}[0, \pi_i^*] \tag{5.16}$$

where $\pi_i^* = \min_{k:z_{ik}=1} \{\pi_{ik}\}$.

Sampling $\boldsymbol{z}_{ik}$

We let $z_i^1 = z_{i,z_{ik}=1}, z_i^0 = z_{i,z_{ik}=0}$, the likelihood term becomes:

$$\Pr(\boldsymbol{e}_{ij} | Z_{\setminus i}, z_i^1, W) = \sigma(z_i^1 W z_j)^{\boldsymbol{e}_{ij}} (1 - \sigma(z_i^1 W z_j))^{1-\boldsymbol{e}_{ij}} \tag{5.17}$$

Thus, we get:

$$\begin{aligned} &\Pr(z_{ik} | \pi_i, \{\boldsymbol{e}_{ij}\}_{j=1}^n, Z_{\setminus i}, W) \propto \\ &\begin{cases} \pi_{ik} \prod_j \left[\Pr(\boldsymbol{e}_{ij} | Z_{\setminus i}, z_i^1, W) \Pr(\boldsymbol{e}_{ji} | Z_{\setminus i}, z_i^1, W)\right], & z_{ik} = 1; \\ (1-\pi_{ik}) \prod_j \left[\Pr(\boldsymbol{e}_{ij} | Z_{\setminus i}, z_i^0, W) \Pr(\boldsymbol{e}_{ji} | Z_{\setminus i}, z_i^0, W)\right], & z_{ik} = 0. \end{cases} \end{aligned} \tag{5.18}$$

Sampling W_{kl}

Due to the nonconjugacy of $\sigma(\cdot)$ function, we use the Metropolis-Hastings method to do the sampling. Setting the proposal distribution the same as

the prior distribution Normal$(0, \sigma_W)$, we have the acceptance ratio as:

$$A(W_{kl}^*, W_{kl}^\tau) = \min\left\{1, \frac{f(W_{kl}^*)}{f(W_{kl}^\tau)}\right\}. \tag{5.19}$$

Sampling Hyper-parameters λ_f, $\boldsymbol{\mu}_f$, λ_v, λ_W

For $\boldsymbol{\mu}_f$, we set the prior as Gaussian prior Normal$(0, \lambda_\mu)$, which leads to:

$$\begin{aligned} p(\boldsymbol{\mu}_f|\lambda_\mu, \boldsymbol{\eta}, \lambda_f) &\propto \text{Normal}(\boldsymbol{\mu}_f; 0, \lambda_\mu) \prod_k \text{Normal}(\boldsymbol{\eta}_{fk}; \boldsymbol{\mu}_f, \lambda_f) \\ \propto \quad \text{Normal}&\left(\boldsymbol{\mu}_f; \frac{\sum_k \boldsymbol{\eta}_{fk}}{\lambda_f^2 + K}, 1 + \frac{K}{\lambda_f^2}\right) \end{aligned} \tag{5.20}$$

For the rest of the hyper-parameters, we set the vague gamma prior $\mathcal{G}(a., b.)$ on them and the corresponding update can be done accordingly.

For λ_f, we give the prior on λ_f^{-2}:

$$\begin{aligned} p(\lambda_f|a_f, b_f, \eta, \boldsymbol{\mu}_f) &\propto \mathcal{G}(\lambda_f^{-2}; a_f, b_f) \prod_f \prod_k \text{Normal}(\boldsymbol{\eta}_{fk}; \boldsymbol{\mu}_f, \lambda_f) \\ \propto \quad \mathcal{G}&\left(\lambda_f^{-2}; a_f + \frac{1}{2}KF, b_f + \frac{1}{2}\sum_k \sum_f (\boldsymbol{\eta}_{fk} - \boldsymbol{\mu}_f)^2\right) \end{aligned} \tag{5.21}$$

For λ_v, we give the prior on λ_v^{-2}:

$$\begin{aligned} p(\lambda_v|a_v, b_v, \eta, \phi) &\propto \mathcal{G}(\lambda_v^{-2}; a_v, b_v) \prod_i \text{Normal}(v_{i:}; \boldsymbol{\eta}\phi_{i:}^T, \lambda_v) \\ \propto \quad \mathcal{G}&\left(\lambda_v^{-2}; a_v + \frac{1}{2}KN, b_v + \frac{1}{2}\sum_k \sum_i (v_{ik} - \boldsymbol{\eta}_k\phi_i)^2\right) \end{aligned} \tag{5.22}$$

For λ_W, we give the prior on λ_W^{-2}:

$$\begin{aligned} p(\lambda_W|a_W, b_W, W) &\propto \mathcal{G}(\lambda_W^{-2}; a_W, b_W) \prod_k \prod_l \text{Normal}(W_{kl}; 0, \lambda_W) \\ \propto \quad \text{Gamma}&\left(\lambda_W^{-2}; a_W + \frac{1}{2}K^2, b_W + \frac{1}{2}\sum_k \sum_i W_{kl}^2\right) \end{aligned} \tag{5.23}$$

5.4.3 π_i-Collapsed Sampling for the niMM Model

When the community number is know in advance, inferencing the niMM model by collapsing the mixed-membership distributions $\{\pi_i\}_i^n$ is a promising solution. W.l.o.g., the membership indicators' joint probability for node i is:

$$\Pr(\{s_{ij}\}_{j=1}^n, \{r_{ji}\}_{j=1}^n | \phi, \eta) \propto \frac{\prod_k \text{Gamma}(\text{Gamma}(N_{ik} + \prod_f \eta_{fk}^{\phi_{if}}))}{\text{Gamma}(2n + \sum_k \prod_f \eta_{fk}^{\phi_{if}})} \tag{5.24}$$

$\forall k \in \{1, \cdots, K\}$, the conditional probability of the membership indicator s_{ij} (the same to r_{ij}) is:

$$\Pr(s_{ij} = k | \{s_{ij_0}\}_{j_0 \neq j}, \{r_{ji}\}_{j_0=1}^n, \phi, \eta) \propto N_{ik}^{\setminus s_{ij}} + \prod_f \eta_{fk}^{\phi_{if}} \tag{5.25}$$

Compared to its counterpart in MMSB:

$$\Pr(s_{ij} = k | \{s_{ij_0}\}_{j_0 \neq j}, \{r_{ji}\}_{j_0=1}^n, \alpha, K) \propto N_{ik}^{\setminus s_{ij}} + \frac{\alpha}{K} \tag{5.26}$$

our collapsed niMM model (cniMM) replaces the term of $\frac{\alpha}{K}$ in Eq. (5.26) with $\{\prod_f \eta_{fk}^{\phi_{if}}\}_{k=1}^K$. In fact, while the MMSB generates the mixed-membership distribution π_i through the Dirichlet distribution with parameters $(\frac{\alpha}{K}, \cdots, \frac{\alpha}{K})$, our cniMM's corresponding one is the Dirichlet distribution with unequal parameter $(\prod_f \eta_{f1}^{\phi_{if}}, \cdots, \prod_f \eta_{fK}^{\phi_{if}})$.

Due to the unknown information on the undiscovered communities, we limit our cniMM model into this finite communities' number case. The extension on the infinite communities' case remains an interesting future task.

5.4.4 Computational Complexity

We estimate the computational complexities for each model and present the results in Table 5.1. Our niMM and niLF are $\mathcal{O}(Kn^2 + Kn + FKn)$ and $\mathcal{O}(K^2n^2 + Kn + FKn)$ respectively, with $\mathcal{O}(Kn)$ for the sampling of $\{\pi_i\}_{i=1}^n$ and $\mathcal{O}(FKn)$ for the incorporation of node information.

Table 5.1: Computational Complexity for Different Models

Models	Computational complexity
IRM	$\mathcal{O}(K^2L)$[1]((Palla et al. 2012))
LFRM	$\mathcal{O}(K^2n^2)$ ((Palla et al. 2012))
MMSB	$\mathcal{O}(Kn^2)$ ((Kim et al. 2012))
NMDR	$\mathcal{O}(Kn^2 + Kn + FKn) = \mathcal{O}(Kn^2)$
niMM	$\mathcal{O}(Kn^2 + Kn + FKn) = \mathcal{O}(Kn^2)$
niLF	$\mathcal{O}(K^2n^2 + Kn + FKn) = \mathcal{O}(K^2n^2)$

[1] $L = \sum_{i,j} \boldsymbol{e}_{ij}$, denotes the number of positive link data.

5.5 Experiments

We analyze the performance of our models (niMM and niLF) on two real-world datasets: the Lazega dataset (Lazega 2001) and the MIT Reality Mining dataset (Eagle & Sandy 2006). The comparison models we are using include IRM (Kemp et al. 2006), LFRM (Miller et al. 2009), iMMM (Koutsourelakis & Eliassi-Rad 2008) (an infinite community case of MMSB (Airoldi et al. 2008)), and NMDR (Kim et al. 2012).

We have independently implemented the above baseline models to the best of our understanding. There has been a slight variation to NMDR, in which we have employed Gibbs sampling to sample the unknown cluster number, instead of the Retrospective MCMC (Papaspiliopoulos & Roberts 2008) used in the original chapter. This setting is to ensure a fair comparison as all of our sampling schemes are under the Gibbs sampling pipeline.

To validate our models' link prediction performance, we use a ten-folds cross-validation strategy. For each node's link data, we randomly select one out of ten from them as the test data. Then, we remove these test data and keep the remaining ones as the training data. The corresponding evaluation criteria (Kim et al. 2012) are the training error ($0-1$ loss) on the training data, the testing error ($0-1$ loss), the testing log likelihood and the AUC

Table 5.2: Performance on Lazega Dataset (Mean $\mp$ Standard Deviation)

Models	Training error	Testing error	Testing log likelihood	AUC
IRM	0.099 ∓ 0.000	0.105 ∓ 0.001	-201.791 ∓ 3.350	0.706 ∓ 0.017
LFRM	0.057 ∓ 0.002	0.105 ∓ 0.006	-222.592 ∓ 16.199	0.817 ∓ 0.020
iMMM	0.049 ∓ 0.007	0.110 ∓ 0.003	-202.715 ∓ 5.308	0.807 ∓ 0.014
NMDR	0.064 ∓ 0.006	0.113 ∓ 0.002	-207.719 ∓ 3.475	0.829 ∓ 0.011
niMM	$\mathbf{0.033 \mp 0.006}$	0.107 ∓ 0.002	$\mathbf{-196.050 \mp 4.396}$	$\mathbf{0.837 \mp 0.012}$
niLF	0.039 ∓ 0.013	$\mathbf{0.101 \mp 0.003}$	-213.525 ∓ 12.325	0.812 ∓ 0.014
cniMM	0.047 ∓ 0.009	0.112 ∓ 0.002	-205.067 ∓ 4.532	0.831 ∓ 0.012

Table 5.3: Performance on Reality Dataset (Mean $\mp$ Standard Deviation)

Models	Training error	Testing error	Testing log likelihood	AUC
IRM	0.063 ∓ 0.000	0.067 ∓ 0.000	-133.804 ∓ 1.127	0.826 ∓ 0.005
LFRM	0.040 ∓ 0.002	0.063 ∓ 0.004	-143.607 ∓ 10.059	0.853 ∓ 0.018
iMMM	0.030 ∓ 0.006	0.063 ∓ 0.002	$\mathbf{-126.788 \mp 3.478}$	0.862 ∓ 0.012
NMDR	0.039 ∓ 0.004	0.067 ∓ 0.001	-139.523 ∓ 2.937	0.857 ∓ 0.014
niMM	$\mathbf{0.027 \mp 0.005}$	$\mathbf{0.062 \mp 0.002}$	-127.738 ∓ 3.131	0.851 ∓ 0.013
niLF	0.038 ∓ 0.005	0.073 ∓ 0.005	-131.037 ∓ 9.452	$\mathbf{0.865 \mp 0.014}$
cniMM	0.055 ∓ 0.002	0.064 ∓ 0.001	-126.909 ∓ 2.646	0.860 ∓ 0.010

(Area Under the roc Curve) score on the test data. Apart from this, we also conduct a study on learning the node information's importance indicator in the Lazega dataset.

At the beginning of the learning process, we set the vague Gamma prior Gamma$(1, 1)$ for the hyper-parameters $\alpha_\eta, \beta_\eta, \alpha_W, \beta_W$. For W's setting, we set Beta$(1, 1)$ as the prior distribution. For the attributes values that are not in binary form, we have to do the binary transform. The initial states are of random guesses on the hidden labels (membership indicators in MMSB and latent feature in LFRM). For all the experiments, we run chains of $10,000$ MCMC samples 30 times, assuming the first 5000 samples are used for burn-in. The average statistics of the remaining 5000 samples are reported.

5.5.1 Performance on the Lazega Dataset

The Lazega dataset includes the social network links within a US firm in 1988 to 1991. The dataset contains a co-work network for 71 attorneys, in which each directional link data is labelled as 1 (exist) or 0 (absent). Apart from this 71 binary asymmetric matrices, the dataset also provides information on each node (i.e. attorneys), including the Status (partner or associate), Gender, Office (Boston, Hartford or Providence), Years (with the firm), Age, Practice (litigation or corporate), and Law school (harvard, yale, ucon or other). After binarizing these attributes, we obtain a 71×11 binary information matrix.

Link prediction by different models is conducted and the results are shown in Table 5.3. Notably, the performance of our implementation of NMDR model is inferior compared to its original ((Kim et al. 2012)), which may be a result of a sub-optimal metadata binarization process. However, we have shown that, with the same attributes, our niMM model performs better than the NMDR model, as well as other relational models without the involving of node information. The performance of cniMM is also quite competitive.

Attribute-community importance learning

Another interesting issue here is the attribute-community importance learning for the importance indicator $\boldsymbol{\eta}$. As the learning should fix the communities during the sampling iteration, we use cniMM to observe the effect of node attribute information on each individual community. The number of communities is set to 4 and the results are shown in Table 5.4. Each value shows the geometric mean of the $5,000$ samples in the MCMC inference. Also, we should note smaller value indicates larger influence. For notational clarity, we use the black bold symbol to denote the values that under 0.5.

Table 5.4: Attribute-community importance learning for $\boldsymbol{\eta}$

Community		1	2	3	4
Office	boston	**0.3103**	1.3139	**0.0877**	2.7415
	hartford	**0.4061**	0.6547	**0.2601**	0.9010
Age	young	1.1884	1.0649	0.8954	1.2016
	middle	0.8562	0.7420	0.7078	0.9639
Years	long	**0.3684**	0.5422	**0.2089**	1.8316
	middle	0.7429	0.7164	0.6534	1.3045
School	yale	0.9733	0.6881	0.9465	0.7372
	ucon	1.4117	1.0636	1.1856	0.8408
Status	partner	0.9822	0.8203	0.9583	0.6359
Practice	litigation	**0.2971**	0.9731	**0.3405**	0.9884
Gender	man	**0.3972**	1.1592	0.7156	0.8653

As we can see, the importance of attributes office in boston and hartford, long years with firm and litigation in practice is the smallest amongst all attributes. This implies they are more important than others in affecting hidden community formation. This is generally consistent with our commonsense. For instance, people in the same office would usually have more communications in everyday life; employees would be more familiar with each other if they together have a longer time stay with the firm. The result of

Table 5.5: Mixing rate (Mean $\mp$ Standard Deviation) for different models, with the bold type denoting the best ones within each row.

Datasets	Lazega		Reality	
Criteria	$\hat{\tau}$	ESS	$\hat{\tau}$	ESS
iMMM	166.2 ∓ 90.37	77.6 ∓ 38.71	184.9 ∓ 78.88	62.5 ∓ 22.70
LFRM	310.6 ∓ 141.95	40.7 ∓ 26.26	113.4 ∓ 77.35	125.5 ∓ 71.93
NMDR	179.8 ∓ 156.96	134.3 ∓ 133.12	142.8 ∓ 129.99	185.0 ∓ 206.12
niMM	$\mathbf{39.1 \mp 40.58}$	$\mathbf{341.8 \mp 132.00}$	$\mathbf{27.8 \mp 22.49}$	$\mathbf{449.7 \mp 181.37}$
niLF	149.2 ∓ 126.12	61.2 ∓ 59.93	134.2 ∓ 163.23	71.24 ∓ 48.74

the importance of the litigation in practice seems a bit interesting probably because the litigation needs frequent corporations and thus leads to the connection.

5.5.2 MIT Reality Mining

Based on the MIT Reality Mining dataset (Eagle & Sandy 2006), we obtain a proximity matrix describing each node's proximity towards the others, i.e., e_{ij} represents the proximity from i to j based on participant i's opinion. With the same setting of the previous model (Koutsourelakis & Eliassi-Rad 2008), we manually set the proximity value to be larger than 10 minutes per day as 1, and 0 otherwise. We hence obtain a 73×73 asymmetric matrix.

Alongside this directional link data, we also have survey data on the participants' information (i.e. node information), including the transport choice to work, social activity, the communication method, and satisfaction of university life. As we can see in Table 5.3, we find our niMM and niLF models' performances are competitive in relation to the ones in iMMM, however, we do not achieve a significant improvement compared to the baseline models. When we trace back to the node information, we find it does not have a direct correlations with the link data. This may be the main reason for our models' less significant result.

5.5.3 Convergence Behaviour

Trace plot for AUC

A trace plot for the AUC value versus iteration time could help us choose an appropriate burn-in length. An earlier reach to the stable status of MCMC is desirable as it indicates fast convergence. Figure 5.2 and Figure 5.3 show the detailed results.

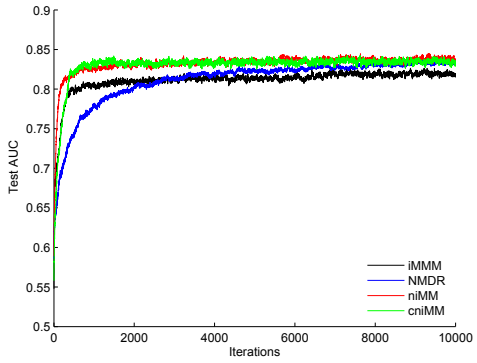

Figure 5.2: Trace plot of the AUC value versus iteration time in the Lazega dataset

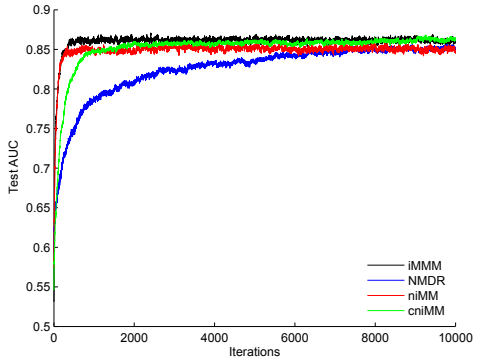

Figure 5.3: Trace plot of the AUC value versus iteration time the Reality dataset

As we can see, except for NMDR, all the other models reach the stable status quite fast. On the Lazega dataset, our niMM and cniMM outper-

form all the others. On the MIT Reality dataset, our niMM and cniMM's performances are still quite competitive.

Mixing rate for a stable MCMC.

In addition to the MCMC trace plot, another interesting observation is the mixing rate of the stable MCMC chains. We use the number of active communities K as a function of the updated variable to monitor the mixing rate of the MCMC samples, whereas the efficiency of the algorithms can be measured by estimating the integrated autocorrelation time τ and Effective Sample Size (ESS) for K. τ is a good performance indicator as it measures the statistical error of Monte Carlo approximation on a target function f. The smaller the τ, the more efficient the algorithm. Also, the ESS of the stable MCMC chains informs the quality of the Markov chains, i.e., a larger ESS value indicates more independent useful samples, which is our desired property.

On estimating the integrated autocorrelation time, different approaches are proposed in (Geyer 1992). Here we use an estimator $\widehat{\tau}$ (Kalli et al. 2011) and the ESS value is calculated based on $\widehat{\tau}$ as:

$$\widehat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \widehat{\rho}_l;\ ESS = \frac{2M}{1+\widehat{\tau}}. \tag{5.27}$$

Here $\widehat{\rho}_l$ is the estimated autocorrelation at lag l and C is a cut-off point which is defined as $C := \min\{l : |\widehat{\rho}_l| < 2/\sqrt{M}\}$, and M is equal to half of the original sample size, as the first half is treated as a burn in phase. The detailed results are shown in Table 5.5. As we can see, our model niMM performs the best among all the models.

5.6 Summary

Increasing applications with natural and social networking behaviors request the effective modeling of hidden relations and structures. This is beyond the currently available models, which only involve limited link information in

binary settings. In this chapter, we have proposed a unified transform to incorporate the rich node attribute information into the relational models. The proposed node-information involved mixed membership (niMM) model and node-information involved latent feature (niLF) model have been demonstrated to be effective in learning the structure and have shown advanced performance on learning implicit relations and structures.

In the stick-breaking construction of $C1$ in Section 5.3.1, we could put an additional α into the product of $\prod_f \eta_{fk}^{\phi_{if}}$. This combined result $\alpha \prod_f \eta_{fk}^{\phi_{if}}$ can avoid the concentration parameter to be 1 in case all of the $\{\phi_{if}\}_f$ are 0. While the conjugate property can still be kept, this also set an extra parameter for the model. In a result, additional computational complexity is required for this incorporation.

5.7 Limitations & Future Work

In this chapter, the main limitation may be the fixed "linear" combinations of different nodes' information. The nodes' information should be combined with different manners, as well as different weights. Another issue is the incorporation method into the communities, which should be more flexible.

On the future work, the work here is expected to investigate the following: 1) how to integrate the multi-relational networks and unify them into the niMM framework to deeply understand network structures; 2) as there are more advanced constructions for the beta process (Paisley et al. 2010)(Paisley, Blei & Jordan 2012), what are more flexible ways to incorporate the node information into LFRM; and 3) when the node information goes beyond the binary scope and becomes the continuous form, how can this information been utilized.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, various strategies on the modelling of coupling relations in the relational data have been presented. After an introduction of the preliminary knowledge and literature review, the coupling relations have been individually described in Chapters $3, 4, 5$, which include the coupling relations between the mixed-membership distributions across time, the coupling relations within the communities of the networks and the coupling relations between the node information and the mixed-membership distribution.

In Chapter 2, preliminary knowledge on the nonparametric Bayesian methods, the Monte Carlo Markov Chain (MCMC) methods and the two representative work of mixed-membership stochastic blockmodel and latent feature relational model has been given. Also, extensive literature reviews on the coupling relation learning, relational models and stick-breaking process. All of which worth a closer look in comparing to the works in this thesis have been given individually.

In Chapter 3, a generalised and flexible framework to further improve the popular mixed-membership stochastic blockmodel by allowing a network to have infinite types of communities with relationships that change across time have been provided. This is noted to be as describing the coupling

relations between the mixed-membership distributions across time. By incorporating a time-sticky factor for the mixed-membership distributions, the time-correlation between latent labels have been realistically modelled. Both Gibbs sampling and adapted Slice-Efficient sampling have been used to inference the desired target distribution. Quantity analysis on the MCMC's convergence behavior, including the convergence test and autocorrelation function, have been provided to enhance the inference performance. The results in the experiments verify that the proposed DIM3 from this study is effective for learning the coupling relation between the mixed-membership distribution across time.

In Chapter 4, a new framework to capture the coupling relations within the communities of the networks has been proposed, which is to greatly complement the mixed-membership stochastic blockmodel which essentially treats nodes independently. The principal contribution of the proposed model is the introduction of the Copula function into MMSB, which represents the correlation between the pair of membership indicators, while keeping the membership indicators' marginal distribution invariant. The results show that, using both synthetic and real data, the Copula-incorporated MMSB, i.e., cMMSB, is effective in learning the coupling relations within the communities. In terms of inference, the main contribution of this study lies in an analytical solution to both of the conditional marginal likelihoods to the two indicator variables (s_{ij}, r_{ji}), given either the indicator distributions π_i, π_j or the bivariate Copula variables u_{ij}, v_{ij}.

In Chapter 5, a unified approach to incorporate the rich node information into the relational models has been put forward, which is able to describe the coupling relations between the node attribute information and the mixed-membership distribution. The proposed node-information involved mixed membership (niMM) model and node-information involved latent feature (niLF) model have been demonstrated to be effective in learning coupling relations between the node information and mixed-membership distribution and have shown advanced performance on learning implicit relations and

structures.

Each chapter (i.e. from Chapter 3 to Chapter 5) of this thesis is supported by one published paper and others under reviewing as [1] listed in **List of Publications**. Therefore, what has been done and proposed in this thesis is of significance to the coupling relation learning research and the data analytics area.

6.2 Future Work

6.2.1 Future Work on Large-scale Bayesian inference for nonparmetric Bayesian methods

The coupling relations cover far more than what has been done in this thesis. In the future research, more complete coupling relations will be explored, as well as their efficient and effective inference.

In the current age of Big Data, the huge quantity of available data has been exceeding the computational resources available. Thus, there is increasing interest for new large-scale learning methods, including the Bayesian inference methods for the nonparametric Bayesian priors. The Bayesian inference methods provide complete characterizations of the joint posterior distribution over the model parameter and hidden variable, rather than seeking an optimal point estimate in an optimization manner (either loss function or maximum likelihood estimation). In this way, the Bayesian inference methods are better at modelling the uncertainty, as well as avoiding the over-fitting problem with the introduction of prior information (Teh, Thiéry & Vollmer 2014).

The detail large-scale Bayesian inference methods of the nonparametric Bayesian prior can be roughly categorized into several parts. 1), stochastic variational inference methods (Hoffman et al. 2013), which applies the s-

[1] The paper of chapter 3 is published, and the paper of chapter 4,5 are still under review under journal or conference

tochastic optimization techniques to optimize the variational bound (such as KL divergence); 2), submodular variational inference (Reed & Ghahramani 2013), which is to use the submodular maximization techniques in getting a lower variational bound. Currently, this work is applied in the Indian Buffet Process prior only; 3), parallel MCMC sampling (Williamson, Dubey & Xing 2013)(Neiswanger, Wang & Xing 2013), which attempts distribute the whole inference computation to several parts; 4), subsampling MCMC methods (Welling & Teh 2011)(Ahn, Balan & Welling 2012)(Ahn, Shahbaba & Welling 2014), which is trying to update the solver by a small subset of the data.

6.2.2 Future work from the relational models perspective

However, several issues still exist in using the relational models to describe the coupling relations. In this case, a few will be listed here.

(i). **To what extent does relational modelling address the coupling relations?** The coupling relations cover the complex interactions between the objects, while the relational models only consider the simplest interactions of these, which is mainly represented as the binary (or real value) interaction. Also, this dependency relation mainly focuses on the inter-node relation, while it does little in the intra-node relation.

(ii). **What aspects of the coupling relations have not been covered by relational models ?** Currently, the relational models only address the following simplest cases:

- relational models only consider two-nodes interaction, which is an over-simplified assumption. In real-world application, the interaction usually occurs among more than two nodes. For instance, in

the stock-trading market, the buy and sell options happen among different levels of people.

- relational models use the communities compatibilities to determine the interaction value, which is also set in a vacuum. In a real-world application, communities influence in the interaction usually plays as a small role in some cases, and in other cases, the interaction is affected by various other information, including the nodes profile, the specific environment (this may be related to time dynamic, space information, or even cultural custom), or even the goal (in the stock market, making profit is the goal and the relations may move towards this goal).

(iii). **Where are the main gaps between coupling relations and the relational models?** The main gaps between the coupling relations and the relational models mainly lie in the perfect assumption in the relational models, which is seldom seen in real world application. From another perspective, this has also stimulated several opportunities for working on the coupling relation learning part. More specifically, the following options are available for further research.

- considering the interactions between more nodes. In this case, there may be a need to incorporate the matrix or even the tensor to represent the relations.

- incorporating more information in building the relations, including the nodes profile, time or space information and goal.

- instead of considering the inter-node dependence, in coupling relation learning, there are more forms in representing the relations, such as their intra-node relation, time correlation. Utilizing this correlations to depict the whole picture would be an interesting future task.

6.2.3 Future work on the inconsistent estimators of the component number

(Miller & Harrison 2013) has shown that even a simple DP mixture model estimator is inconsistent for the number of mixture components. In all of our models of this thesis, we feel it as a missing part since we did not pay much attention to this specific topic. The number of mixture components discovered here is mainly used for the MCMC convergence diagnose. However, we should emphasize that we did encounter the situations where there are more smaller-size clusters than the reality, especially in the work of *nonparametric power-law data clustering*. This can be remained as a future work and I believe this topic would be both interesting and influential for the future research.

Appendix A

Derivation equations for the model of MTV-g

In the section, we derive the equations for the MCMC sampling in MTV-g in chapter 3. The variables we want to sample are: $\boldsymbol{\beta}$, $\boldsymbol{Z} = \{s_{ij}^t, r_{ij}^t\}_{n\times n}^{1:T}$ and $\hat{\boldsymbol{m}}$.

A.1 Sample $\boldsymbol{\beta}$

$\boldsymbol{\beta}$ represents each component's proportion.

$$(\beta_1, \cdots, \beta_K, \beta_\mu) \sim Dir(\hat{m}_{\cdot 1}, \cdots, \hat{m}_{\cdot K}, \gamma) \tag{A.1}$$

Here $\hat{m}_{\cdot k}$ denotes the dish k's whole *considered* table. (We use dish and community alternatively, however, they stand for the same meaning.)

A.2 Sample $Z = \{s_{ij}^t, r_{ij}^t\}_{n\times n}^{1:T}$

$\forall i, j \in \{1, \cdots, n\}, t \in \{1, \cdots, T\}$, we sequentially sample the pair (s_{ij}^t, r_{ij}^t) together, as they jointly determine $\boldsymbol{e}_{ij}^t$'s indexed position (i.e. row and col-

umn) in W.

$$\begin{aligned} & P(s_{ij}^t = k, r_{ij}^t = l | \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \boldsymbol{e}, \boldsymbol{\beta}, \alpha, \lambda_1, \lambda_2, \kappa) \\ \propto & P(s_{ij}^t = k, r_{ij}^t = l | \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \boldsymbol{\beta}, \alpha, \kappa) \\ & \cdot P(\boldsymbol{e}_{ij}^t | \boldsymbol{e} \backslash \{\boldsymbol{e}_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2) \end{aligned} \tag{A.2}$$

The first term of Eq. (A.2), i.e., the prior of (s_{ij}^t, r_{ij}^t) is:

$$\begin{aligned} & \Pr(s_{ij}^t = k, r_{ij}^t = l | \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \boldsymbol{\beta}, \alpha, \kappa) \\ \propto & \Pr(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \\ & \cdot \Pr(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0 \neq i}, \{s_{j i_0}\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}) \\ & \cdot \prod_{l=1}^{2n} \Pr(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}) \\ & \cdot \prod_{l=1}^{2n} \Pr(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}) \end{aligned} \tag{A.3}$$

We treat the first part of Eq. (A.3) as:

$$
\begin{aligned}
&\Pr(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \\
\propto &\Pr(s_{ij}^t = k, \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \\
= &\int_{\boldsymbol{\pi}_i^t} p(\boldsymbol{\pi}_i^t | \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \cdot \Pr(s_{ij}^t = k | \boldsymbol{\pi}_i^t) \prod_{j_0 \neq j} \Pr(s_{ij_0}^t | \boldsymbol{\pi}_i^t) \prod_{j_0=1}^n \Pr(r_{j_0 i}^t | \boldsymbol{\pi}_i^t) d\boldsymbol{\pi}_i^t \\
= &\int_{\boldsymbol{\pi}_i^t} \frac{\Gamma(\sum_l \alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1})}{\prod_l \Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1})} \prod_l \boldsymbol{\pi}_{il}^{\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1} - 1} \prod_l \boldsymbol{\pi}_{il}^{N_{il}^{t,-s_{ij}^t}} \cdot \boldsymbol{\pi}_{ik} d\boldsymbol{\pi}_i^t \\
= &\frac{\Gamma(\sum_l \alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1})}{\prod_l \Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1})} \frac{\prod_l \Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1} + N_{il}^{t,-s_{ij}^t}) + \delta(l,k)}{\Gamma(\sum_l \alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1} + N_{il}^{t,-s_{ij}^t} + \delta(l,k))} \\
= &\frac{\Gamma(\alpha + 2n \cdot \kappa)}{\Gamma(\alpha + 2n \cdot \kappa + 2n)} \prod_l \frac{\Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1} + N_{il}^{t,-s_{ij}^t} + \delta(l,k))}{\Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1})} \\
= &\frac{\Gamma(\alpha + 2n \cdot \kappa)}{\Gamma(\alpha + 2n \cdot \kappa + 2n)} \prod_l \frac{\Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1} + N_{il}^{t,-s_{ij}^t})}{\Gamma(\alpha \boldsymbol{\beta}_l + \kappa N_{il}^{t-1})} \\
&\cdot (\alpha \boldsymbol{\beta}_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t}) \\
\propto &\alpha \boldsymbol{\beta}_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t}
\end{aligned}
\tag{A.4}
$$

Here $N_{ik}^t = \sum_{j_0=1}^n \mathbf{1}(s_{ij_0}^t = k) + \sum_{j_0=1}^n \mathbf{1}(r_{j_0 i}^t = k)$, $N_{i\cdot}^t = \sum_k N_{ik}^t = 2n$; $N_{ik}^{t,-s_{ij}^t} = \sum_{j_0 \neq j}^n \mathbf{1}(s_{ij_0}^t = k) + \sum_{j_0=1}^n \mathbf{1}(r_{j_0 i}^t = k)$, $N_{i\cdot}^{t,-s_{ij}^t} = \sum_k N_{ik}^t - 1 = 2n - 1$; $\delta(l,k)$ is the dirac delta function with $\delta(l,k) = \begin{cases} 1, & l = k; \\ 0, & l \neq k. \end{cases}$

Thus, we get

$$
\begin{aligned}
&\Pr(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}) \\
\propto &\begin{cases} \alpha \boldsymbol{\beta}_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t}, & k \in \{1, \cdots, K\}; \\ \alpha \boldsymbol{\beta}_\mu, & k = K+1. \end{cases}
\end{aligned}
\tag{A.5}
$$

Under the similar transformation, we got

$$
\begin{aligned}
&\Pr(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0 \neq i}, \{s_{ji_0}\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}) \\
\propto &\begin{cases} \alpha \boldsymbol{\beta}_l + \kappa N_{jl}^{t-1} + N_{jl}^{t,-r_{ij}^t}, & l \in \{1, \cdots, K\}; \\ \alpha \boldsymbol{\beta}_\mu, & l = K+1. \end{cases}
\end{aligned}
\tag{A.6}
$$

Similarly, we got:

$$\begin{aligned}&\prod_{l=1}^{2n}\Pr(z_{il}^{t+1}|z_{i.}^{t}/s_{ij}^{t},s_{ij}^{t}=k,\boldsymbol{\beta},\alpha,\kappa,N_{i}^{t+1})\\&\propto\frac{\Gamma(\alpha\beta_{k}+N_{ik}^{t+1}+\kappa N_{ik}^{t,-s_{ij}^{t}}+\kappa)}{\Gamma(\alpha\beta_{k}+N_{ik}^{t+1}+\kappa N_{ik}^{t,-s_{ij}^{t}})}\frac{\Gamma(\alpha\beta_{k}+\kappa N_{ik}^{t,-s_{ij}^{t}})}{\Gamma(\alpha\beta_{k}+\kappa N_{ik}^{t,-s_{ij}^{t}}+\kappa)}\end{aligned}\tag{A.7}$$

$$\begin{aligned}&\prod_{l=1}^{2n}\Pr(z_{jl}^{t+1}|z_{j.}^{t}/r_{ij}^{t},r_{ij}^{t}=l,\boldsymbol{\beta},\alpha,\kappa,N_{j}^{t+1})\\&\propto\frac{\Gamma(\alpha\boldsymbol{\beta}_{l}+N_{jl}^{t+1}+\kappa N_{jl}^{t,-r_{ij}^{t}}+\kappa)}{\Gamma(\alpha\boldsymbol{\beta}_{l}+N_{jl}^{t+1}+\kappa N_{jl}^{t,-r_{ij}^{t}})}\cdot\frac{\Gamma(\alpha\boldsymbol{\beta}_{l}+\kappa N_{jl}^{t,-r_{ij}^{t}})}{\Gamma(\alpha\boldsymbol{\beta}_{l}+\kappa N_{jl}^{t,-r_{ij}^{t}}+\kappa)}\end{aligned}\tag{A.8}$$

Also, the second term of Eq. (A.2), i.e., the likelihood of (s_{ij}^{t},r_{ij}^{t}) becomes as:

$$\begin{aligned}&P(\boldsymbol{e}_{ij}^{t}|\boldsymbol{e}\backslash\{\boldsymbol{e}_{ij}^{t}\},s_{ij}^{t}=k,r_{ij}^{t}=l,\boldsymbol{Z}\backslash\{s_{ij}^{t},r_{ij}^{t}\},\lambda_{1},\lambda_{2})\\&\propto P(\boldsymbol{e}_{ij}^{t},\boldsymbol{e}\backslash\{\boldsymbol{e}_{ij}^{t}\},s_{ij}^{t}=k,r_{ij}^{t}=l,\boldsymbol{Z}\backslash\{s_{ij}^{t},r_{ij}^{t}\},\lambda_{1},\lambda_{2})\\&\propto\int_{\eta_{k,l}}P(\boldsymbol{e}_{ij}^{t}|s_{ij}^{t}=k,r_{ij}^{t}=l,\eta_{k,l})\cdot p(\eta_{k,l}|\lambda_{1},\lambda_{2})\\&\quad\cdot\prod_{s_{i'j'}^{t}=k,r_{i'j'}^{t}=l,i'j'\neq ij}P(e_{i'j'}^{t}|\eta_{k,l})d\eta_{k,l}\\&\propto\int_{\eta_{k,l}}\eta_{k,l}^{\boldsymbol{e}_{ij}^{t}+n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^{t}}+\lambda_{1}-1}(1-\eta_{k,l})^{1-\boldsymbol{e}_{ij}^{t}+n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^{t}}+\lambda_{2}-1}d\eta_{k,l}\\&\propto\frac{\Gamma(\boldsymbol{e}_{ij}^{t}+n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^{t}}+\lambda_{1})\Gamma(1-\boldsymbol{e}_{ij}^{t}+n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^{t}}+\lambda_{2})}{\Gamma(1+n_{k,l}^{t,-\boldsymbol{e}_{ij}^{t}}+\lambda_{1}+\lambda_{2})}\end{aligned}\tag{A.9}$$

Here $n_{k,l}^{t,-\boldsymbol{e}_{ij}^{t}}=n_{k,l}^{t}-\mathbf{1}(s_{ij}^{t}=k,r_{ij}^{t}=l)=\sum_{i'j'}\mathbf{1}(s_{i'j'}^{t}=k,r_{i'j'}^{t}=l)-\mathbf{1}(s_{ij}^{t}=k,r_{ij}^{t}=l)$, $n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^{t}}=n_{k,l}^{1,t}-\mathbf{1}(s_{ij}^{t}=k,r_{ij}^{t}=l)\boldsymbol{e}_{ij}^{t}=\sum_{s_{i'j'}^{t}=k,r_{i'j'}^{t}=l}e_{i'j'}^{t}-\mathbf{1}(s_{ij}^{t}=k,r_{ij}^{t}=l)\boldsymbol{e}_{ij}^{t}$ and $n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^{t}}=n_{k,l}^{t,-\boldsymbol{e}_{ij}^{t}}-n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^{t}}$.

Considering the case of $\boldsymbol{e}_{ij}^{t}=1$ and $\boldsymbol{e}_{ij}^{t}=0$, Eq. (A.9) is to be simplified

as:

$$P(\boldsymbol{e}_{ij}^t|\boldsymbol{e}\backslash\{\boldsymbol{e}_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z}\backslash\{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2)$$
$$= \begin{cases} \frac{n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} + \lambda_1}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & e_{ij}^t = 1; \\ \frac{n_{k,l}^{t,0,-e_{ij}^t} + \lambda_2}{n_{k,l}^{t,-e_{ij}^t} + \lambda_1 + \lambda_2}, & e_{ij}^t = 0. \end{cases} \tag{A.10}$$

By combing the results of Eq. (A.5)(A.6)(A.7)(A.8)(A.10), the sampling of (s_{ij}^t, r_{ij}^t) is completed.

A.3 Sampling $\hat{m}$

$\hat{m}$'s sampling is detailed in the chapter.

Appendix B

Several fundamental distributions used in the thesis

B.1 Bernoulli distribution

The probability mass function of the Bernoulli distribution is:

$$\forall 0 \leq p \leq 1, \Pr(x) = \begin{cases} p, & x = 1; \\ 1 - p, & x = 0. \end{cases} \tag{B.1}$$

B.2 Multinomial distribution

The probability mass function of the Multinomial distribution is:

$$\forall \sum_{k=1}^{K} p_k = 1, \Pr(n_1, \cdots, n_K) = \frac{n!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} p_k^{n_k}. \tag{B.2}$$

Here $\forall k \in \{1, \cdots, K\}, n_k$ denotes the number of times the k^{th} component appears, and p_k represents the probability of the k^{th} component appears.

B.3 Beta distribution

The probability density function of the Beta distribution is:

$$\forall \alpha, \beta > 0, p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}. \tag{B.3}$$

Here $\Gamma(\cdot)$ is the Gamma function, which is $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.

B.4 Dirichlet distribution

The probability density function of the Dirichlet distribution is:

$$\forall \alpha_1, \cdots, \alpha_k > 0, p(x_1, \cdots, x_K) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k - 1}. \tag{B.4}$$

Here $(x_1, \cdots, x_K)$ lies in a K-dimensional simplex, which is $\sum_{k=1}^K x_k = 1$.

B.5 Gamma distribution

The probability density function of the Gamma distribution is:

$$\forall \alpha, \beta > 0, p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}. \tag{B.5}$$

Here α is the shape parameter, while β is the rate parameter.

Appendix C

List of Symbols

The following list is neither exhaustive nor exclusive, but may be helpful.

n	number of nodes
K	number of discovered communities
T	number of whole time stamps
t	the specific time stamp
e_{ij}^t	directional, binary interactions at time t
$\boldsymbol{\beta}$	a stick-breaking representation to denote the "significance" of all existing communities at all times
γ, α	concentration parameters for HDP
κ	a sticky parameter representing the time-persistence effect
s_{ij}^t	sender's (from i to j) membership indicator at time t
r_{ij}^t	receiver's (from j to i) membership indicator at time t
Z	all the membership indicators, i.e. $Z = \{s_{ij}^t, r_{ij}^t\}_{i,j,t}$

$z_{i\cdot}^t$	node i's membership indicators at time t, i.e. $\{s_{ij}^t, r_{ji}^t\}_{j=1}^n$
$\boldsymbol{m}_{ik}^t$	in Chinese Restaurant Franchise analogy, the number of tables eating dish k for restaurant i at time t
π_i^t	mixed-membership distribution for node i at time t, it generates $s_{i1}^t, \cdots, s_{in}^t, r_{1i}^t, \cdots, r_{ni}^t$
π_{ik}^t	the "significance" of community k for node i at time t
W	role-compatibility matrix
$W_{k,l}$	compatibilities between communities k and l
$n_{k,l}^t$	number of links from community k to l at time t i.e. $n_{k,l}^t = \#\{ij : s_{ij}^t = k, r_{ij}^t = l.\}$
$n_{k,l}^{t,1}$	part of $m_{k,l}$ where the corresponding $e_{ij}^t = 1$ at time t, i.e. $n_{k,l}^{t,1} = \sum_{s_{ij}^t=k, r_{ij}^t=l} e_{ij}^t$
$n_{k,l}^{t,0}$	part of $m_{k,l}$ where the corresponding $e_{ij}^t = 0$ at time t, i.e. $n_{k,l}^{t,0} = n_{k,l}^t - n_{k,l}^{t,1}$
N_{ik}^t	number of times that a node i has participated in community k (either sending or receiving) at time t, i.e. $N_{ik}^t = \#\{j : s_{ij}^t = k\} + \#\{j : r_{ji}^t = k\}$
e_{ij}	directional, binary interactions
s_{ij}	sender's (from i to j) membership indicator
r_{ij}	receiver's (from j to i) membership indicator
π_i	mixed-membership distribution for node i, it generates $s_{i1}, \cdots, s_{in}, r_{1i}, \cdots, r_{ni}$
π_{ik}	the "significance" of community k for node i

$m_{k,l}$	number of links from community k to l, i.e. $m_{ik} = \#\{ij : s_{ij} = k, r_{ij} = l.\}$
$m^1_{k,l}$	part of $m_{k,l}$ where the corresponding $e_{ij} = 1$, i.e. $m^1_{k,l} = \sum_{s_{ij}=k,r_{ij}=l} e_{ij}$
$m^0_{k,l}$	part of $m_{k,l}$ where the corresponding $e_{ij} = 0, m^0_{k,l} = m_{k,l} - m^1_{k,l}$
θ	parameter associated with any Copula function
F	number of attributes in node information
ϕ	an $n \times F$ binary matrix, $\phi_{if} = 1$ denotes the i^{th} data occupies the f^{th} attribute
η	an $F \times K$ positive matrix, η_{fk} indicates the importance of f^{th} attribute to k^{th} roles.
ψ_i	stick-breaking weights to constitute π_i
z_i	latent feature vector of node i in LFRM
N_{ik}	number of times that a node i has participated in community k (either sending or receiving), i.e. $N_{ik} = \#\{j : s_{\boldsymbol{i}j} = k\} + \#\{j : r_{j\boldsymbol{i}} = k\}$

Bibliography

Ahn, S., Balan, A. K. & Welling, M. (2012), Bayesian posterior sampling via stochastic gradient fisher scoring, *in* 'Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012', pp. 1591–1598.

Ahn, S., Shahbaba, B. & Welling, M. (2014), Distributed stochastic gradient MCMC, *in* 'Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014', pp. 1044–1052.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008), 'Mixed membership stochastic blockmodels', *The Journal of Machine Learning Research* **9**, 1981–2014.

Aldous, D. J. (1985), *Exchangeability and related topics*, Springer.

Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. (2003), 'An introduction to mcmc for machine learning', *Machine learning* **50**(1-2), 5–43.

Antoniak, C. (1974), 'Mixtures of dirichlet processes with applications to bayesian nonparametric problems', *The Annals of Statistics* pp. 1152–1174.

Asur, S., Parthasarathy, S. & Ucar, D. (2009), 'An event-based framework for characterizing the evolutionary behavior of interaction graphs', *ACM Trans. Knowl. Discov. Data* **3**(4), 16:1–16:36.

Bishop, C. M. et al. (2006), *Pattern Recognition and Machine Learning*, Vol. 1, springer New York.

Blei, D. M., Griffiths, T. L. & Jordan, M. I. (2010), 'The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies', *Journal of ACM* **57**(2), 7:1–7:30.

Bouguila, N. & Ziou, D. (2010), 'A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling', *IEEE Transactions on Neural Networks* **21**(1), 107–122.

Breiger, R. L., Boorman, S. A. & Arabie, P. (1975), 'An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling', *Journal of Mathematical Psychology* **12**(3), 328–383.

Cao, L. (2010), 'Domain-driven data mining: Challenges and prospects', *Knowledge and Data Engineering, IEEE Transactions on* **22**(6), 755–769.

Cao, L. (2013), 'Combined mining: Analyzing object and pattern relations for discovering and constructing complex yet actionable patterns', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **3**(2), 140–155.

Cao, L. (2014), 'Non-iidness learning in behavioral and social data', *Comput. J.* **57**(9), 1358–1370.

Cao, L., Dai, R. & Zhou, M. (2009), 'Metasynthesis: M-space, m-interaction, and m-computing for open complex giant systems', *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* **39**(5), 1007–1021.

Cao, L., Luo, D. & Zhang, C. (2009), Ubiquitous intelligence in agent mining, *in* 'Agents and Data Mining Interaction', Springer, pp. 23–35.

Cao, L., Ou, Y. & Yu, P. S. (2012), 'Coupled behavior analysis with applications', *Knowledge and Data Engineering, IEEE Transactions on* **24**(8), 1378–1392.

Cao, L., Ou, Y., Yu, P. S. & Wei, G. (2010), Detecting abnormal coupled sequences and sequence changes in group-based manipulative trading behaviors, *in* 'Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 85–94.

Cao, L., Zhang, C. & Zhou, M. (2008), 'Engineering open complex agent systems: A case study', *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **38**(4), 483–496.

Caron, F., Davy, M. & Doucet, A. (2007), Generalized polya urn for time-varying dirichlet process mixtures, *in* 'Uncertainty in Artificial Intelligence', pp. 33–40.

Chazottes, J.-R., Collet, P., Külske, C. & Redig, F. (2007), 'Concentration inequalities for random fields via coupling', *Probability Theory and Related Fields* **137**(1-2), 201–225.

Chen, C., Ding, N. & Buntine, W. L. (2012), Dependent hierarchical normalized random measures for dynamic topic modeling, *in* 'Proceedings of the 30th International Conference on Machine Learning (ICML-2012)', pp. 895–902.

Chen, C., Rao, V., Buntine, W. L. & Teh, Y. W. (2013), Dependent normalized random measures, *in* 'Proceedings of the 30th International Conference on Machine Learning (ICML-2013)', pp. 969–977.

Cheng, X., Miao, D., Wang, C. & Cao, L. (2013), Coupled term-term relation analysis for document clustering, *in* 'Neural Networks (IJCNN), The 2013 International Joint Conference on', IEEE, pp. 1–8.

Chung, Y. & Dunson, D. B. (2011), 'The local dirichlet process', *Annals of the Institute of Statistical Mathematics* **63**(1), 59–80.

Clauset, A., Moore, C. & Newman, M. E. (2008), 'Hierarchical structure and the prediction of missing links in networks', *Nature* **453**(7191), 98–101.

Coleman, J. S. et al. (1964), 'Introduction to mathematical sociology', *London Free Press Glencoe.* .

Dunson, D. B. (2006), 'Bayesian dynamic modeling of latent trait distributions', *Biostatistics* **7**(4), 551–568.

Dunson, D. B. & Park, J.-H. (2008), 'Kernel stick-breaking processes', *Biometrika* **95**(2), 307–323.

Eagle, N. & Sandy, A. (2006), 'Reality mining: sensing complex social systems', *Personal Ubiquitous Comput.* **10**(4), 255–268.

Ferguson, T. S. (1973), 'A bayesian analysis of some nonparametric problems', *The annals of statistics* pp. 209–230.

Fortunato, S. (2010), 'Community detection in graphs', *Physics Reports* **486**(3), 75–174.

Foti, N. J., Futoma, J. D., Rockmore, D. N. & Williamson, S. (2013), A unifying representation for a class of dependent random measures, *in* 'Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, (AISTATS-2013)', pp. 20–28.

Foulds, J. R., DuBois, C., Asuncion, A. U., Butts, C. T. & Smyth, P. (2011), A dynamic relational infinite feature model for longitudinal social networks, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 287–295.

Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. (2008), An hdp-hmm for systems with state persistence, *in* 'Proceedings of the 25th International Conference on Machine Learning', ICML '08, ACM, New York, NY, USA, pp. 312–319.

Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. (2011*a*), 'A sticky hdp-hmm with application to speaker diarization', *The Annals of Applied Statistics* **5**(2A), 1020–1056.

Fox, E., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. (2011*b*), 'Bayesian nonparametric inference of switching dynamic linear models', *Signal Processing, IEEE Transactions on* **59**(4), 1569–1585.

Freeman, S. C. & Freeman, L. C. (1979), *The networkers network: A study of the impact of a new communications medium on sociometric structure*, School of Social Sciences University of Calif.

Fu, W., Song, L. & Xing, E. (2009), Dynamic mixed membership blockmodel for evolving networks, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ACM, pp. 329–336.

Gallager, R. (2009), 'Discrete stochastic processes', *Rice University* .

Gelman, A. & Rubin, D. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* **7**(4), 457–472.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *in* 'Bayesian Statistics', University Press, pp. 169–193.

Geyer, C. J. (1992), 'Practical markov chain monte carlo', *Statistical Science* **7**(4), 473–483.

Girvan, M. & Newman, M. (2002), 'Community structure in social and biological networks', *Proceedings of the National Academy of Sciences* **99**(12), 7821–7826.

Gopalan, P., Gerrish, S., Freedman, M., Blei, D. M. & Mimno, D. M. (2012), Scalable inference of overlapping communities, *in* 'Advances in Neural Information Processing Systems', pp. 2249–2257.

Griffiths, T. L. & Ghahramani, Z. (2005), Infinite latent feature models and the indian buffet process, *in* 'Advances in Neural Information Processing Systems 18', MIT Press, pp. 475–482.

Griffiths, T. L. & Ghahramani, Z. (2011), 'The indian buffet process: An introduction and review', *Journal of Machine Learning Research* **12**, 1185–1224.

Guo, Z.-C. & Shi, L. (2011), 'Classification with non-iid sampling', *Mathematical and Computer Modelling* **54**(5), 1347–1364.

Heaukulani, C. & Ghahramani, Z. (2013), Dynamic probabilistic models for latent feature propagation in social networks, *in* 'Proceedings of the 30th International Conference on Machine Learning (ICML-13)', pp. 275–283.

Heidelberger, P. & Welch, P. D. (1981), 'A spectral method for confidence interval generation and run length control in simulations', *Commun. ACM* **24**(4), 233–245.

Herlau, T., Morup, M., Schmidt, M. N. & Hansen, L. K. (2012), Detecting hierarchical structure in networks, *in* 'Cognitive Information Processing (CIP), 2012 3rd International Workshop on', IEEE, pp. 1–6.

Hjort, N., Holmes, C., Müller, P. & Walker, S. (2010), *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Ho, Q., Parikh, A. P. & Xing, E. P. (2012), 'A multiscale community blockmodel for network exploration', *Journal of the American Statistical Association* **107**(499), 916–934.

Ho, Q., Song, L. & Xing, E. P. (2011), Evolving cluster mixed-membership blockmodel for time-evolving networks, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 342–350.

Hoff, P. D. (2009), 'Multiplicative latent factor models for description and prediction of social networks', *Computational and Mathematical Organization Theory* **15**(4), 261–272.

Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. (2013), 'Stochastic variational inference', *The Journal of Machine Learning Research* **14**(1), 1303–1347.

Hofman, J. M. & Wiggins, C. H. (2008), 'Bayesian approach to network modularity', *Physical review letters* **100**(25), 258701.

Hunter, D. R., Goodreau, S. M. & Handcock, M. S. (2008), 'Goodness of fit of social network models', *Journal of the American Statistical Association* **103**(481).

Isella, L., Stehl, J., Barrat, A., Cattuto, C., Pinton, J. & Van den Broeck, W. (2011), 'What's in a crowd? analysis of face-to-face behavioral networks', *Journal of Theoretical Biology* **271**(1), 166–180.

Ishiguro, K., Iwata, T., Ueda, N. & Tenenbaum, J. B. (2010), Dynamic infinite relational model for time-varying relational data analysis., *in* 'NIPS', Curran Associates, Inc., pp. 919–927.

Ishiguro, K., Ueda, N. & Sawada, H. (2012), Subset infinite relational models, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 547–555.

Ishwaran, H. & James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453).

Jebara, T. (2004), Maximum entropy discrimination, *in* 'Machine Learning', Springer, pp. 61–98.

Kalli, M., Griffin, J. & Walker, S. (2011), 'Slice sampling mixture models', *Statistics and Computing* **21**(1), 93–105.

Kapferer, B. (1972), *Strategy and transaction in an African factory: African workers and Indian management in a Zambian town*, Manchester University Press.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. & Ueda, N. (2006), Learning systems of concepts with an infinite relational model, *in* 'Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1', AAAI'06, AAAI Press, pp. 381–388.

Kim, D. I., Hughes, M. & Sudderth, E. B. (2012), The nonparametric metadata dependent relational model, *in* J. Langford & J. Pineau, eds, 'Proceedings of the 29th International Conference on Machine Learning (ICML-12)', ACM, New York, NY, USA, pp. 1559–1566.

Kim, D. I. & Sudderth, E. B. (2011), The doubly correlated nonparametric topic model, *in* 'Advances in Neural Information Processing Systems 24', Curran Associates, Inc., pp. 1980–1988.

Klimt, B. & Yang, Y. (2004), The enron corpus: A new dataset for email classification research, *in* 'Machine Learning: ECML 2004', Springer, pp. 217–226.

Knowles, D. & Ghahramani, Z. (2007), Infinite sparse factor analysis and infinite independent components analysis, *in* 'Independent Component Analysis and Signal Separation', Springer, pp. 381–388.

Koutsourelakis, P. & Eliassi-Rad, T. (2008), Finding mixed-memberships in social networks, *in* 'Proceedings of the 2008 AAAI Spring Symposium on Social Information Processing', pp. 48–53.

Lazega, E. (2001), *The Collegial Phenomenon: The Social Mechanisms of Co-operation Among Peers in a Corporate Law Partnership*, Oxford University Press on Demand.

Li, J., Wang, C., Cao, L. & Yu, P. S. (n.d.), Efficient selection of globally optimal rules on large imbalanced data based on rule coverage relationship

analysis, *in* 'Proceedings of the 2013 SIAM International Conference on Data Mining', pp. 216–224.

Lin, D. & Fisher, J. (2012), Coupling nonparametric mixtures via latent dirichlet processes, *in* 'Advances in Neural Information Processing Systems 25', pp. 55–63.

Lin, D., Grimson, E. & Fisher, J. W. (2010), Construction of dependent dirichlet processes based on poisson processes, *in* 'Advances in Neural Information Processing Systems 23', Curran Associates, Inc., pp. 1396–1404.

Lin, Y.-R., Chi, Y., Zhu, S., Sundaram, H. & Tseng, B. L. (2009), 'Analyzing communities and their evolutions in dynamic social networks', *ACM Trans. Knowl. Discov. Data* **3**(2), 8:1–8:31.

Liu, J. S. (1994), 'The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem', *Journal of the American Statistical Association* **89**(427), 958–966.

MacEachern, S. N. (1999), Dependent nonparametric processes, *in* 'ASA proceedings of the section on bayesian statistical science', American Statistical Association, pp. 50–55, Alexandria, VA, pp. 50–55.

McNeil, A. J. & Nešlehová, J. (2009), 'Multivariate archimedean copulas, d-monotone functions and l_1-norm symmetric distributions', *The Annals of Statistics* pp. 3059–3097.

Miller, J. W. & Harrison, M. T. (2013), A simple example of dirichlet process mixture inconsistency for the number of components, *in* 'Advances in Neural Information Processing Systems', pp. 199–206.

Miller, K., Jordan, M. I. & Griffiths, T. L. (2009), Nonparametric latent feature models for link prediction, *in* 'Advances in Neural Information Processing Systems', pp. 1276–1284.

Mohri, M. & Rostamizadeh, A. (2009), Rademacher complexity bounds for non-iid processes, *in* 'Advances in Neural Information Processing Systems', pp. 1097–1104.

Mohri, M. & Rostamizadeh, A. (2010), 'Stability bounds for stationary φ-mixing and β-mixing processes', *The Journal of Machine Learning Research* **11**, 789–814.

Mørup, M. & Schmidt, M. (2012), 'Bayesian community detection', *Neural computation* **24**(9), 2434–2456.

Morup, M., Schmidt, M. N. & Hansen, L. K. (2011), Infinite multiple membership relational modeling for complex networks, *in* 'Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on', IEEE, pp. 1–6.

Müller, P. & Quintana, F. A. (2004), 'Nonparametric bayesian data analysis', *Statistical science* pp. 95–110.

Neiswanger, W., Wang, C. & Xing, E. (2013), 'Asymptotically exact, embarrassingly parallel mcmc', *arXiv preprint arXiv:1311.4780* .

Nelsen, R. (2006), *An introduction to copulas*, Springer.

Newcomb, T. M. (1961), *The acquaintance process*, Holt, Rinehart & Winston.

Newman, M. E. & Girvan, M. (2004), 'Finding and evaluating community structure in networks', *Physical review E* **69**(2), 026113.

Nowicki, K. & Snijders, T. (2001), 'Estimation and prediction for stochastic blockstructures', *Journal of the American Statistical Association* **96**(455), 1077–1087.

Paisley, J. W., Blei, D. M. & Jordan, M. I. (2012), Stick-breaking beta processes and the poisson process, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 850–858.

Paisley, J., Wang, C. & Blei, D. M. (2012), 'The discrete infinite logistic normal distribution', *Bayesian Analysis* **7**(4), 997–1034.

Paisley, J., Zaas, A., Woods, C. W., Ginsburg, G. S. & Carin, L. (2010), A stick-breaking construction of the beta process, *in* 'International Conference on Machine Learning', pp. 847–854.

Palla, K., Ghahramani, Z. & Knowles, D. A. (2012), An infinite latent attribute model for network data, *in* 'Proceedings of the 29th International Conference on Machine Learning (ICML-12)', ACM, pp. 1607–1614.

Papaspiliopoulos, O. & Roberts, G. (2008), 'Retrospective markov chain monte carlo methods for dirichlet process hierarchical models', *Biometrika* **95**(1), 169–186.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006), 'Coda: Convergence diagnosis and output analysis for mcmc', *R News* **6**(1), 7–11.

Rao, V. & Teh, Y. W. (2009), Spatial normalized gamma processes, *in* 'Advances in Neural Information Processing Systems', pp. 1554–1562.

Rasmussen, C. E. (1999), 'The infinite gaussian mixture model.', **12**, 554–560.

Reed, C. & Ghahramani, Z. (2013), Scaling the indian buffet process via submodular maximization, *in* 'Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013', pp. 1013–1021.

Ren, L., Du, L., Carin, L. & Dunson, D. (2011), 'Logistic stick-breaking process', *The Journal of Machine Learning Research* **12**, 203–239.

Rodriguez, A. & Dunson, D. B. (2011), 'Nonparametric bayesian models through probit stick-breaking processes', *Bayesian Analysis* **6**(1), 145–177.

Roy, D. M., Kemp, C., Mansinghka, V. & Tenenbaum, J. B. (2007), 'Learning annotated hierarchies from relational data'.

Roy, D. M. & Teh, Y. W. (2009), The mondrian process, *in* 'Advances in Neural Information Processing Systems', pp. 1377–1384.

Sampson, S. F. (1969), 'Crisis in a cloister', *doctoral dissertation in department of psychology, Cornell University* .

Sarkar, P. & Moore, A. W. (2005), 'Dynamic social network analysis using latent space models', *ACM SIGKDD Explorations Newsletter* **7**(2), 31–40.

Sarkar, P., Siddiqi, S. M. & Gordon, G. J. (2007), A latent space approach to dynamic embedding of co-occurrence data, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 420–427.

Schmidt, M. & Morup, M. (2013), 'Nonparametric bayesian modeling of complex networks: an introduction', *Signal Processing Magazine, IEEE* **30**(3), 110–128.

Schmidt, M. N., Herlau, T. & Mørup, M. (2013), 'Nonparametric bayesian models of hierarchical structure in complex networks', *arXiv preprint arXiv:1311.1033* .

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Sethuraman, J. (1994), 'A constructive definition of dirichlet priors', *Statistica Sinica* **4**, 639–650.

Shachter, R. D. (1998), Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams), *in* 'Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 480–487.

Tang, L. & Liu, H. (2010), 'Community detection and mining in social media', *Synthesis Lectures on Data Mining and Knowledge Discovery* **2**(1), 1–137.

Teh, Y. W., Görür, D. & Ghahramani, Z. (2007), Stick-breaking construction for the indian buffet process, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 556–563.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006), 'Hierarchical Dirichlet processes', *Journal of the American Statistical Association* **101**(476), 1566–1581.

Teh, Y. W., Thiéry, A. & Vollmer, S. (2014), 'Consistency and fluctuations for stochastic gradient langevin dynamics', *arXiv preprint arXiv:1409.0578* .

Thibaux, R. & Jordan, M. I. (2007), Hierarchical beta processes and the indian buffet process, *in* 'International conference on Artificial Intelligence and Statistics', pp. 564–571.

Van Gael, J., Saatci, Y., Teh, Y. & Ghahramani, Z. (2008), Beam sampling for the infinite hidden markov model, *in* 'Proceedings of the 25th International Conference on Machine Learning', ACM, pp. 1088–1095.

Walker, S. (2007), 'Sampling the dirichlet mixture model with slices', *Communications in Statistics Simulation and Computation®* **36**(1), 45–54.

Walker, S. G., Damien, P., Laud, P. W. & Smith, A. F. (1999), 'Bayesian nonparametric inference for random distributions and related functions', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 485–527.

Wang, C. & Cao, L. (2012), Modeling and analysis of social activity process, *in* 'Behavior Computing', Springer, pp. 21–35.

Wang, C., Cao, L., Wang, M., Li, J., Wei, W. & Ou, Y. (2011), Coupled nominal similarity in unsupervised learning, *in* 'Proceedings of the 20th ACM international conference on Information and knowledge management', ACM, pp. 973–978.

Wang, C., She, Z. & Cao, L. (2013*a*), Coupled attribute analysis on numerical data, *in* 'Proceedings of the Twenty-Third international joint conference on Artificial Intelligence', AAAI Press, pp. 1736–1742.

Wang, C., She, Z. & Cao, L. (2013*b*), Coupled clustering ensemble: Incorporating coupling relationships both between base clusterings and objects, *in* 'Data Engineering (ICDE), 2013 IEEE 29th International Conference on', IEEE, pp. 374–385.

Wang, Y. & Carin, L. (2012), Levy measure decompositions for the beta and gamma processes, *in* J. Langford & J. Pineau, eds, 'Proceedings of the 29th International Conference on Machine Learning (ICML-12)', ACM, New York, NY, USA, pp. 73–80.

Welling, M. & Teh, Y. W. (2011), Bayesian learning via stochastic gradient langevin dynamics, *in* 'Proceedings of the 28th International Conference on Machine Learning (ICML-11)', pp. 681–688.

Williamson, S., Dubey, A. & Xing, E. P. (2013), Parallel markov chain monte carlo for nonparametric mixture models, *in* 'Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013', pp. 98–106.

Xing, E., Fu, W. & Song, L. (2010), 'A state-space mixed membership blockmodel for dynamic network tomography', *The Annals of Applied Statistics* **4**(2), 535–566.

Xu, Z., Tresp, V., Yu, K. & Kriegel, H.-P. (2006), 'Learning infinite hidden relational models', *Uncertainity in Artificial Intelligence (UAI2006)* .

Yang, T., Chi, Y., Zhu, S., Gong, Y. & Jin, R. (2011), 'Detecting communities and their evolutions in dynamic social networks - a bayesian approach', *Machine Learning* **82**(2), 157–189.

Yin, J., Ho, Q. & Xing, E. (2013), A scalable approach to probabilistic latent space inference of large-scale networks, *in* 'Advances in Neural Information Processing Systems', pp. 422–430.

Yu, Y., Wang, C., Gao, Y., Cao, L. & Chen, X. (2013), A coupled clustering approach for items recommendation, *in* 'Advances in Knowledge Discovery and Data Mining', Springer, pp. 365–376.

Zhang, X., Song, L., Gretton, A. & Smola, A. J. (2009), Kernel measures of independence for non-iid data, *in* D. Koller, D. Schuurmans, Y. Bengio & L. Bottou, eds, 'Advances in Neural Information Processing Systems 21', Curran Associates, Inc., pp. 1937–1944.

Zhu, J. (2012), Max-margin nonparametric latent feature models for link prediction, *in* 'Proceedings of the 29th International Conference on Machine Learning (ICML-12)', ACM, pp. 719–726.