Relevance-dependent Biclustering Models for Relational Data Analysis

(関係データ分析のための関連度依存型共クラスタリングモデルの研究)

Iku Ohama

February 2018

Division of Computer Science and Information Technology

Graduate School of Information Science and Technology

Hokkaido University

Abstract

Relational data encoding pairwise relationships between objects appears in many fields. For example point-of-sale (POS) data of an e-commerce (EC) site contain relational data between customers and items, and follower lists in social networking services (SNS) such as Twitter is relational data among users. Recently, with the rapid advancements in internet technologies, a large amount of relational data has been accumulated in many business fields. Therefore, extracting insights by analysing relational data becomes an important challenge for many business persons to refine their business activities.

Biclustering is one of the most popular techniques to extract useful insights from relational data. Biclustering abstracts the given data matrix into a low-dimensional block structure by simultaneously clustering both the row and column objects. For extracting robust bicluster structure from noisy real-world relational data, there have been studied many statistical models for biclustering. Among these models, the Infinite Relational Model (IRM) proposed by Kemp *et al.* is one of the most fundamental biclustering models. The IRM abstracts given relational data into a block structure, in which each block has its own link probability. The IRM can automatically estimates the optimal number of clusters. Furthermore, posterior inference for the IRM can be performed efficiently using collapsed Gibbs sampler.

The IRM and its extended models commonly assume that each block of the bicluster structure has an uniform density. However, this assumption is not acceptable in many

real-world situations. For example, when analyzing shopping behavior of customers on an EC site, there may be items that a certain customer would like to purchase but cannot afford, whereas a customer with a larger budget can purchase any item on sale. Therefore, the bicluster structure underlying real-world relationships may represent highly distorted relationships rather than ideal relationships. To obtain an essential bicluster structure in such a situation, it is natural to assume additional latent factors that affect the observed link probabilities of objects regardless of their membership to the cluster.

In this thesis, in order to overcome the drawback of existing standard biclustering models, we study a novel biclustering problem termed *relevance-dependent biclustering*. In our relevance-dependent biclustering, we assume that each object has an additional latent variable indicating a *relevance* value that determines how strongly the object relates to the cluster. Therefore, the relevance-dependent biclustering can capture block structure with un-uniform density, where the un-uniformness is explained by the effect of the latent relevance values. This is a major advantage of relevance-dependent biclustering because the meaning of obtained clusters can be understood easily by inspecting only a few highly relevant objects.

In Chapter 3, we discuss the relevance-dependency modeling using Boolean functions. More specifically, for each entry of given relational data, we introduce a latent binary variable that indicates whether the entry relates to the block structure (foreground distribution) or to a background noise (background distribution). Then, we introduce a mechanism that the binary variable for an entry is determined by calculating an arbitrary Boolean function of Bernoulli trials from row and column objects. Thus, a probability for the Bernoulli trial can be interpreted as the relevance parameter for the object. By incorporating the mechanism, we propose an extension of the IRM termed the Relevance-Dependent IRM (RDIRM). The RDIRM is also an instance of Bayesian nonparametric models. Therefore, the RDIRM can automatically estimate

the number of clusters. Furthermore, thanks to the conjugacy between component distributions, posterior inference for the RDIRM can be performed efficiently using collapsed Gibbs sampler.

In Chapter 4, we generalize the relevance modeling in the RDIRM. By considering continuous relaxation of the Boolean function in the RDIRM, we propose a mixed-membership mechanism that contains all the Boolean functions as special cases. In the mixed-membership approach, we can resolve two critical limitations of relevance modeling in the RDIRM. First, in the mixed-membership mechanism, the form of the relaxed Boolean function can be automatically estimated from given data. Furthermore, in the mixed-membership mechanism, we can straightforwardly consider relevance values with three or more dimensions. Therefore, we can introduce multiple background distribution for considering different types of irrelevant objects. By incorporating the mixed-membership mechanism, we propose an extension of the RDIRM termed Multi-Layered IRM (MLIRM), which has two background distributions. The relevance parameters in the MLIRM can explain, not only passive objects with few links, but also spamming objects with extremely many links. Similar to the RDIRM, the posterior inference for the MLIRM can also be performed using collapsed Gibbs sampler.

In Chapter 5, we introduce a link function approach for modeling relevance-dependency. More specifically, we introduce the Relevance-dependent Bernoulli Distribution (R-BD), which is a novel prior distribution for relevance-dependent binary matrices. In our R-BD, a link strength for an entry is defined by three non-negative parameters: a typical link strength common to all entries in the matrix, and two relevance parameters for each row and column objects. Then, an observed link probability is directly calculated by transforming the product of these three non-negative variables into a probability using Bernoulli-Poisson link function. The main advantages of the R-BD is as follows. First, the relevance-modeling in the R-BD do not have to consider any

background distributions. Thus, the number of latent variables to be estimated is significantly smaller than those in the RDIRM and MLIRM. Second, the link probability in the R-BD can be modulated widely from 0.0 to 1.0 without introducing complicated mechanism as in the MLIRM. Thus, the effect of relevance values in the R-BD is more interpretable than that in the RDIRM and MLIRM. Finally, as all the parameters of the R-BD can be completely marginalized out, we do not have to explicitly estimate R-BD's parameters when performing posterior inference. By incorporating the R-BD as a component distribution, we propose a novel biclustering model termed the Relevance-dependent Infinite Biclustering (R-IB). Thanks to the property of the R-BD, the posterior inference for the R-IB can also be performed using a collapsed Gibbs sampler. Furthermore, the R-IB can be inferred faster than not only the RDIRM and MLIRM, but also the original IRM.

Finally, we conclude this thesis and discuss future work in Chapter 6. In this study, we introduced the relevance-dependent biclustering problem. Then, we explored several approach for modeling relevance-dependency and developed new relevance-dependent biclustering models for each approach. Furthermore, for each model, we derived an efficient collapsed Gibbs sampler to perform posterior inference. Through this study, the author succeeded in opening the beginning of relevance-dependent biclustering research. For future work, it is of interest to apply the relevance-modeling of this study to more general machine learning problems such as matrix factorization.

Acknowledgements

First and foremost I would like to thank Professor Hiroki Arimura who supervised me from the beginning of my research activity. I also would like to thank Professor Takuya Kida. He directed my research and gave me plenty of advice. I would not accomplish this work without his generous support.

I would like to indicate my gratefulness to Professor Shinichi Minato and Makoto Haraguchi. I thank them for their patient guidance and charitable comments.

I would like to thank everyone in Information Knowledge Network laboratory, especially Ms. Yu Manabe. Her supports significantly facilitated my research activities.

I would like to express my appreciation to all my current and past colleagues in Panansonic corporation. They gave me an opportunity to start collaborative research with Professor Hiroki Arimura and Takuya Kida. My research activity in Hokkaido University would not accomplish without their understanding and co-operation.

Finally, I would like to voice my thankfulness to my family. My wife, Emi has been extremely supportive of me throughout my research activity. My son, Issa has continually provided the motivation to finish my degree.

Contents

Chapter 1

Introduction

Relational data encoding pairwise relationships between objects appears in many fields. For example, point-of-sale (POS) data of an e-commerce (EC) site contain relational data between customers and items, and follower lists in social networking services (SNS) such as Twitter is relational data among users. Recently, with the rapid advancements in internet technologies, a large amount of relational data has been accumulated in many business fields. Therefore, extracting insights by analysing such relational data becomes an important challenge for many business persons to refine their business activities. For example, someone might want to obtain the following knowledge:

- How should we categorize customers to help us understand their preferences more clearly?
- How many roles are there in our company? Which employees work on similar tasks?

Biclustering [23, 13, 14, 9, 20, 5, 36] is one of the most popular techniques to extract useful insights from relational data. Biclustering abstracts the given data matrix into a low-dimensional block structure by simultaneously clustering both the row and column objects. For example, biclustering of POS data can be used to elucidate bipartite

relationships between particular customers and particular items that sell well. For extracting robust bicluster structure from noisy real-world relational data, there have been proposed many statistical models for biclustering [62, 63, 47, 34, 72, 71]. Among these models, the Stochastic Block Model (SBM) [47] proposed by Nowicki *et al.* is one of the most fundamental biclustering models. The SBM abstracts given relational data into a block structure, in which each block has its own link probability. More recently, Kemp *et al.* have proposed the Infinite Relational Model (IRM) [34], a Bayesian nonparametric extension of the SBM that can automatically estimates the optimal number of clusters from given relational data. Furthermore, posterior inference for the IRM can be performed efficiently using collapsed inference methods [70, 41, 32, 37]. Because the IRM is so popular, many extensions of the IRM have been proposed [1, 43, 57, 24, 58, 44, 56, 30, 19].

The SBM, IRM and its extended models commonly assume that each block of the bicluster structure has an uniform density (Fig. 1.1a). However, this assumption is not acceptable in many real-world situations. For example, when analyzing shopping behavior of customers on an EC site, there may be items that a certain customer would like to purchase but cannot afford, whereas a customer with a larger budget can purchase any item on sale. Therefore, the bicluster structure underlying real-world relationships may represent highly distorted relationships rather than ideal relationships. To obtain an essential bicluster structure in such a situation, it is natural to assume additional latent factors that affect the observed link probabilities of objects regardless of their membership to the cluster. Although the necessity of considering clusters with uneven density has been discussed in social community analysis [4], this problem has not yet been studied directly.

In this thesis, in order to overcome the drawback of existing *standard biclustering* models (Fig. 1.1a), we study a novel biclustering problem termed *relevance-dependent biclustering* (Fig. 1.1b). In our relevance-dependent biclustering, we assume that each

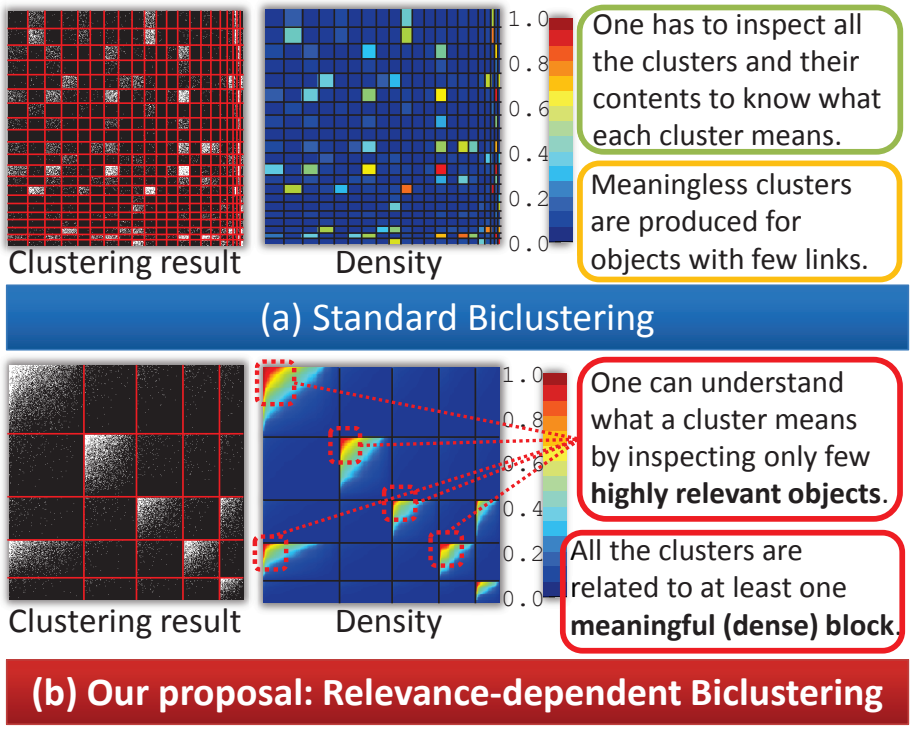


Figure 1.1: Diagrams of (a) standard biclustering and (b) relevance-dependent biclustering.

object has an additional latent variable indicating a *relevance* value that determines how strongly the object relates to the cluster. That is, a large relevance value means that the corresponding object strongly follow the ideal density defined by block structure, whereas a small relevance value indicates that the corresponding object is relatively non-informative and weakly relevant to the block structure. Therefore, the relevance-dependent biclustering can capture block structure with un-uniform densities, where the un-uniformness is explained by the effect of the latent relevance values. This is a major advantage of relevance-dependent biclustering because the meaning of obtained clusters can be understood easily by inspecting only a few highly relevant objects. Furthermore, all obtained clusters are interpretable because they are related to at least one meaningful (dense) block. In this thesis, we explore several approaches

for modeling relevance-dependency in relational data. For each approach, we propose a new relevance-dependent biclustering model that can automatically estimate the number of clusters. Furthermore, we derive an efficient collapsed Gibbs sampling [41] algorithm for performing posterior inference for each proposed model.

In Chapter 3, we discuss the relevance-dependency modeling using Boolean functions. More specifically, for each entry of given relational data, we introduce a latent binary variable that indicates whether the entry relates to the block structure (foreground distribution) or to a background noise (background distribution). Then, we introduce a mechanism that the binary variable for an entry is determined by calculating an arbitrary Boolean function of Bernoulli trials from row and column objects. By introducing probability parameter that controls Bernoulli trial of corresponding object, the probability can be interpreted as the relevance parameter for the object. Thus, a probability for the Bernoulli trial can be interpreted as the relevance parameter for the object. By incorporating the mechanism, we propose an extension of the IRM termed the Relevance-Dependent IRM (RDIRM). The RDIRM is also an instance of Bayesian nonparametric models. Therefore, the RDIRM can automatically estimate the number of clusters. Furthermore, thanks to the conjugacy between component distributions, posterior inference for the RDIRM can be performed efficiently using collapsed Gibbs sampler[1].

In Chapter 4, we generalize the relevance modeling in the RDIRM. By considering continuous relaxation of the Boolean function in the RDIRM, we propose a mixed-membership mechanism that contains all the Boolean functions as special cases. In the mixed-membership approach, we can resolve two critical limitations of relevance modeling in the RDIRM. First, in the mixed-membership mechanism, the form of the relaxed Boolean function can be automatically estimated from given data. Furthermore, in the mixed-membership mechanism, we can straightforwardly consider rele-

[1] This result has been published in [48, 49].

vance values with three or more dimensions. Therefore, we can introduce multiple background distribution for considering different types of irrelevant objects. By incorporating the mixed-membership mechanism, we propose an extension of the RDIRM termed Multi-Layered IRM (MLIRM), which has two background distributions. The relevance parameters in the MLIRM can explain, not only passive objects with few links, but also spamming objects with extremely many links. Similar to the RDIRM, the posterior inference for the MLIRM can also be performed using collapsed Gibbs sampler[2].

In Chapter 5, in order to develop more computationally efficient relevance-dependent biclustering model, we introduce a link function approach for modeling relevance-dependency. More specifically, we introduce the Relevance-dependent Bernoulli Distribution (R-BD), which is a novel prior distribution for relevance-dependent binary matrices. In our R-BD, a link strength for an entry is defined by three non-negative parameters: a typical link strength common to all entries in the matrix, and two relevance parameters for each row and column objects. Then, an observed link probability is directly calculated by transforming the product of these three non-negative variables into a probability using Bernoulli-Poisson link function [74]. The main advantages of the R-BD is as follows. First, the relevance-modeling in the R-BD do not have to consider any background distributions. Thus, the number of latent variables to be estimated is significantly smaller than those in the RDIRM and MLIRM. Second, the link probability in the R-BD can be modulated widely from 0.0 to 1.0 without introducing complicated mechanism as in the MLIRM. Thus, the effect of the relevance values in the R-BD is interpretable. Finally, as the all parameters of the R-BD can be completely marginalized out, we do not have to explicitly estimate R-BD's parameters when performing posterior inference. By incorporating the R-BD as a component distribution, we propose a novel biclustering model termed the Relevance-dependent Infinite Biclus-

[2] This result has been published in [50, 52].

tering (R-IB). Thanks to the property of the R-BD, the posterior inference for the R-IB can also be performed using a collapsed Gibbs sampler. Furthermore, the R-IB can be inferred faster than not only the RDIRM and MLIRM, but also the original IRM[3].

Finally, we conclude this thesis and discuss future work in Chapter 6. In this study, we introduced the relevance-dependent biclustering problem. Then, we explored several approaches for modeling relevance-dependency and developed new relevance-dependent biclustering models for each approach. Furthermore, for each model, we derived an efficient collapsed Gibbs sampler to perform posterior inference. Through this study, the author succeeded in opening the beginning of relevance-dependent biclustering research. For future work, it is of interest to apply the relevance-modeling of this study to more general machine learning problems such as matrix factorization.

[3] This result has been published in [51, 53].

Chapter 2

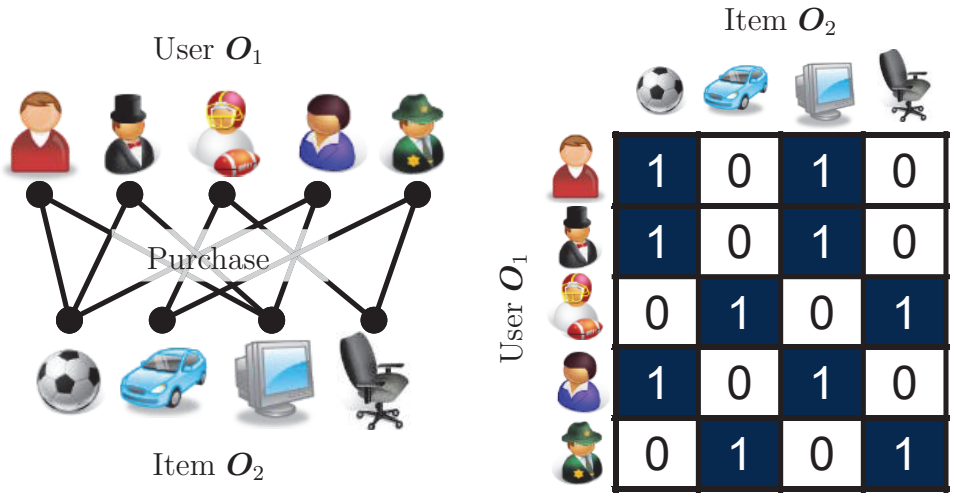
Preliminaries

In this chapter, we introduce basic terms and notations. First, we define the relational data discussed in this thesis. Then, we review the baseline standard biclustering models (i.e., the SBM and IRM). Finally, we discuss the drawbacks of these standard biclustering models.

2.1 Relational Data

Let $\boldsymbol{R}$ be the $I \times J$ binary matrix that represents relational data between a set of objects $T_1 = \{O_{1,i}\}_{i=1}^{I}$ and another set of objects $T_2 = \{O_{2,j}\}_{j=1}^{J}$. An entry $R_{i,j} = 1(0)$ indicates that there is a link (non-link) between $O_{1,i}$ and $O_{2,j}$. For example, customer i's purchase of item j can be represented by $R_{i,j} = 1$. Conversely, $R_{i,j} = 0$ indicates that customer i has not bought item j (Fig. 2.1).

Note that several variations of relational data can be considered straightforwardly. For example, discrete-valued relational data [7, 71, 28, 76, 8, 77, 76, 75] are preferable for encoding customer ratings for items on an EC site. As another example, real-valued relational data [39, 31, 38, 60, 59, 42, 15, 22] can be considered for encoding relationships on sensor networks or traffic volume on transportation networks. In



(a) Graph representation (b) Matrix representation

Figure 2.1: (Best viewed in color.) Diagrams of relational data discussed in this thesis.

addition, we can consider relational data that encode relationships among three or more domains (e.g., customers × items × time) using tensor representation [27, 28]. Furthermore, many other variants of relational data can be considered [64, 40, 18, 67, 68]. Although considering these variants are important, we focus on two-domain binary relationships which is the most common type of the relational data.

2.2 Stochastic Block Model (SBM)

In this section, we briefly review the Stochastic Block Model (SBM) [47], one of the most popular latent variable models for co-clustering relational data. The SBM considers a stochastic distribution over $\boldsymbol{R}$. Let K and L be the number of clusters for T_1 and T_2, respectively. The SBM assumes latent variables $z_{1,i} \in \{1, \ldots, K\}$ and $z_{2,j} \in \{1, \ldots, L\}$ for T_1 and T_2, respectively. These latent variables indicate cluster assignments for objects. That is, $z_{1,i} = k$ means that the i-th row object is assigned to k-th row

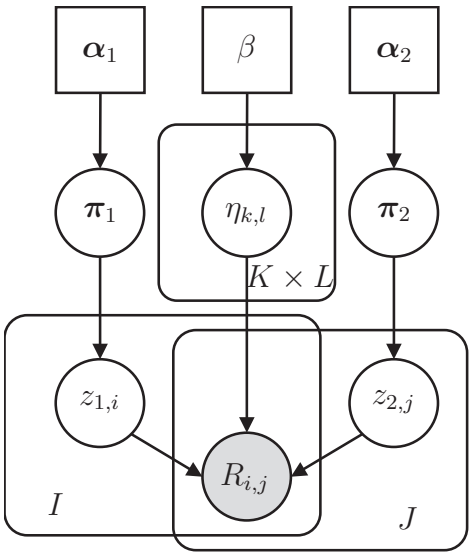


Figure 2.2: Graphical representation for the generative model of the SBM. Circle nodes denote random variables, square nodes denote hyperparameters, shaded nodes denote observations, and round-edged squares indicate number of individual variables. Directed connections denote probabilistic dependencies.

cluster. Similarly, $z_{2,j} = l$ means that the j-th column object is assigned to l-th column cluster. Note that, in this thesis, we also use 1-of-K representation for $z_{1,i}$ and $z_{2,j}$ as $\boldsymbol{Z}_{1,i} = \{Z_{1,i,k}\}_{k=1}^{K} \in \{0,1\}^K$ and $\boldsymbol{Z}_{2,j} = \{Z_{2,j,l}\}_{l=1}^{L} \in \{0,1\}^L$, respectively, where $\sum_{k=1}^{K} Z_{1,i,k} = \sum_{l=1}^{L} Z_{2,j,l} = 1$. Therefore, a situation that i-th row object is assigned to k-th row cluster is described as $z_{1,i} = k$ or $Z_{1,i,k} = 1$ through this theses. Let $\boldsymbol{\eta}$ be the $K \times L$ matrix of link probabilities between K clusters for T_1 and L clusters for T_2, where $\eta_{k,l} \in [0,1]$ indicates the probability that there is a link between an object assigned to cluster k and an object assigned to cluster l. Thus, in the SBM, the link probability between $O_{1,i}$ and $O_{2,j}$ is given as follows:

$$P(R_{i,j} = 1 \,|\, z_{1,i}, z_{2,j}, \boldsymbol{\eta}) = \eta_{z_{1,i}, z_{2,j}}, \tag{2.1}$$

which is also represented as

$$P(R_{i,j} = 1 \mid \boldsymbol{Z}_{1,i}, \boldsymbol{Z}_{2,j}, \boldsymbol{\eta}) = \boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j} = \sum_{k=1}^{K} \sum_{l=1}^{L} Z_{1,i,k} Z_{2,j,l} \eta_{k,l}. \tag{2.2}$$

As we can easily see from the form of Eq. (2.2), the SBM factorizes relational data $\boldsymbol{R}$ into three low dimensional matrices $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, and $\boldsymbol{\eta}$. In the SBM, a Dirichlet distribution is assumed as a prior for each cluster assignment, and a beta distribution is assumed as a prior for a link probability between two clusters $\eta_{k,l}$. More specifically, the full description of the generative model for the SBM is as follows:

$$\boldsymbol{\pi}_1 \mid \boldsymbol{\alpha}_1 \sim \text{Dirichlet}(\boldsymbol{\alpha}_1), \tag{2.3}$$

$$\boldsymbol{\pi}_2 \mid \boldsymbol{\alpha}_2 \sim \text{Dirichlet}(\boldsymbol{\alpha}_2), \tag{2.4}$$

$$\boldsymbol{Z}_{1,i} \mid \boldsymbol{\pi}_1 \sim \text{Categorical}(\boldsymbol{\pi}_1), \tag{2.5}$$

$$\boldsymbol{Z}_{2,j} \mid \boldsymbol{\pi}_2 \sim \text{Categorical}(\boldsymbol{\pi}_2), \tag{2.6}$$

$$\eta_{k,l} \mid \beta \sim \text{Beta}(\beta, \beta), \tag{2.7}$$

$$R_{i,j} \mid \boldsymbol{\eta}, \boldsymbol{Z}_{1,i}, \boldsymbol{Z}_{2,j} \sim \text{Bernoulli}\left(\boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j}\right), \tag{2.8}$$

where Dirichlet$(\cdot)$, Categorical$(\cdot)$, Beta$(\cdot, \cdot)$, and Bernoulli$(\cdot)$ are the Dirichlet, categorical, beta, and Bernoulli distributions, respectively. Figure 2.2 shows a graphical representation of the SBM. We now briefly review the above process. First, categorical parameters for row objects $\boldsymbol{\pi}_1$ and column objects $\boldsymbol{\pi}_2$ are drawn from Dirichlet priors with parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, respectively (Eqs. (2.3) and (2.4)). Note that we consider the K and L-dimensional Dirichlet priors for rows and columns, respectively. Second, the cluster assignments $\boldsymbol{Z}_{1,i}$ and $\boldsymbol{Z}_{2,j}$ are drawn from corresponding categorical distributions (Eqs. (2.5) and (2.6)). Here, each row object is assigned to one of K clusters. Similarly, each column object is assigned to one of L clusters. Third, each link probability $\eta_{k,l}$ between row cluster $k \in \{1, \ldots, K\}$ and column cluster $l \in \{1, \ldots, L\}$ is

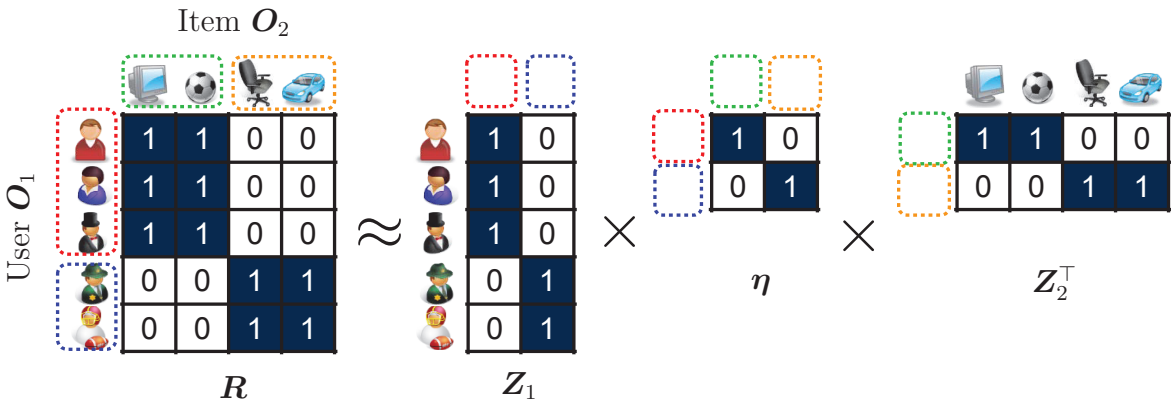


Figure 2.3: (Best viewed in color.) Diagrams of a clustering result obtained by the SBM.

drawn from beta distribution with parameter β (Eq. (2.7)). Finally, a link $R_{i,j}$ between the i-th row object and j-th column object is generated by a Bernoulli trial with probability $\boldsymbol{Z}_{1,i}^\top \boldsymbol{\eta} \boldsymbol{Z}_{2,j}$ (Eq. (2.8)). In the SBM, each object is assigned to one of the finite number of clusters. Therefore, fitting the SBM to given relational data, we can obtain a $K \times L$ block structure. That is why the above model is called the Stochastic Block Model. Figure 2.3 shows the diagram of a clustering result obtained by the SBM.

There are several approaches to perform posterior inference for the SBM. Especially, the Gibbs sampler and the variational Bayes inference are frequently used. The Gibbs sampler guarantees asymptotic convergence to the true posterior by drawing infinitely many samples, whereas the variational Bayes inference is an approximative approach. In this paper, we apply an improved Gibbs sampler called the *collapsed Gibbs sampler* [41] to infer the posterior. In the collapsed Gibbs sampler, some of the model parameters are integrated out. Therefore, we need to sequentially update only the remaining parameters.

Now, we show posteriors for running the collapsed Gibbs sampler for the SBM. For

the SBM, thanks to conjugacy, the Dirichlet parameters $\boldsymbol{\pi}_1, \boldsymbol{\pi}_1$ and the link probabilities $\boldsymbol{\eta}$ can be integrated out. Therefore, the inference of the SBM is performed by sampling only cluster assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. Because $\boldsymbol{z}_2$ can be sampled in the same manner as $\boldsymbol{z}_1$, we concentrate on $\boldsymbol{z}_1$. The conditional posterior for $z_{1,i} = k^*$ is derived as follows:

$$P(z_{1,i} = k^* \mid \boldsymbol{z}_{1,-i}, \boldsymbol{z}_2, \boldsymbol{R}) \propto (m_{1,-i,k^*} + \alpha_{1,k^*}) \times \prod_{l=1}^{L} \frac{B(m^{+i}_{k^*,l} + \beta, \overline{m}^{+i}_{k^*,l} + \beta)}{B(m^{-i}_{k^*,l} + \beta, \overline{m}^{-i}_{k^*,l} + \beta)}, \tag{2.9}$$

where, $\boldsymbol{z}_{1,-i}$ is the cluster assignments for all row objects excluding $O_{1,i}$, $B(\cdot, \cdot)$ is the beta function, and α_{1,k^*} is the k^*-th value of the Dirichlet hyperparameter $\boldsymbol{\alpha}_1$. The symbol $m_{1,-i,k^*}$ denote the number of row objects assigned to cluster k^* excluding $O_{1,i}$. The symbols $m^{+i}_{k^*,l}$, $m^{-i}_{k^*,l}$, $\overline{m}^{+i}_{k^*,l}$, and $\overline{m}^{-i}_{k^*,l}$ denote the numbers of links and non-links, and are computed as follows:

$$m^{+i}_{k^*,l} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^* (z_{1,i} := k^*), z_{2,j} = l}} \sum R_{x,j}, \quad \overline{m}^{+i}_{k^*,l} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^* (z_{1,i} := k^*), z_{2,j} = l}} \sum \overline{R}_{x,j},$$
$$m^{-i}_{k^*,l} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^* (x \neq i), z_{2,j} = l}} \sum R_{x,j}, \quad \overline{m}^{-i}_{k^*,l} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^* (x \neq i), z_{2,j} = l}} \sum \overline{R}_{x,j},$$

where $\overline{R}_{i,j} = 1 - R_{i,j}$. Then, starting from randomly initialized cluster assignments, a collapsed Gibbs solution can be obtained by updating each cluster assignment using Eq. (2.9).

After the burn-in period of Gibbs iterations, the expected a posteriori (EAP) estimation for a link probability $\eta_{k,l}$ that we integrated out can be computed as follows:

$$\eta^{\mathrm{EAP}}_{k,l} = \frac{m_{k,l} + \beta}{m_{k,l} + \overline{m}_{k,l} + 2\beta}, \tag{2.10}$$

where

$$m_{k,l} = \sum_{\substack{i \in T_1, j \in T_2: \\ z_{1,i} = k, z_{2,j} = l}} \sum R_{i,j}, \quad \overline{m}_{k,l} = \sum_{\substack{i \in T_1, j \in T_2: \\ z_{1,i} = k, z_{2,j} = l}} \sum \overline{R}_{i,j}.$$

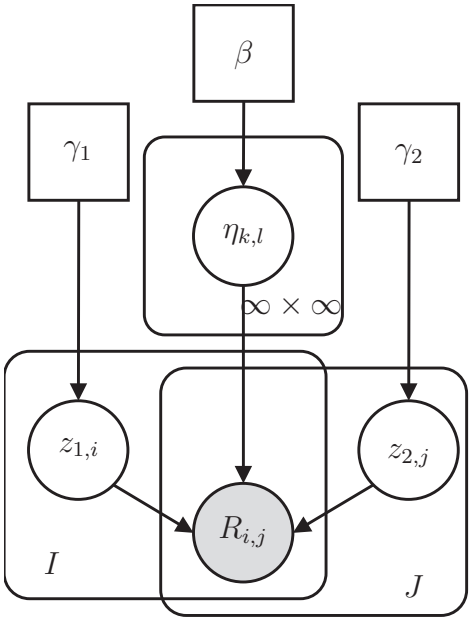


Figure 2.4: Graphical representation for the generative model of the IRM. Circle nodes denote random variables, square nodes denote hyperparameters, shaded nodes denote observations, and round-edged squares indicate number of individual variables. Directed connections denote probabilistic dependencies.

As described above, in the SBM, a prior distribution for cluster assignments is a finite-dimensional Dirichlet distribution. Therefore, we must carefully choose the numbers of clusters K and L to avoid an underfitted or overfitted solution.

2.3 Infinite Relational Model (IRM)

The Infinite Relational Model (IRM) proposed by Kemp *et al.* [34] is a well-known extension of the SBM that can automatically estimate the number of clusters. In the IRM, a Dirichlet Process (DP) [17] is used as a prior distribution for the number of clusters. The DP is a nonparametric stochastic process that can be viewed as an infinite-dimensional Dirichlet distribution. In the IRM, the numbers of mixture

components K and L are theoretically infinite. Therefore, the number of clusters for abstracting relational data is automatically estimated. To implement these infinite mixture models, we can use either a Stick-Breaking Process (SBP) [61] or a Chinese Restaurant Process (CRP) [6, 2]. Typically, a CRP is used to develop a collapsed Gibbs sampler for an infinite mixture model, whereas the SBP is used to develop variational Bayes inference.

The generative process of the IRM with the CRP representation is described as follows:

$$\boldsymbol{Z}_{1,i} \,|\, \gamma_1 \sim \mathrm{CRP}(\gamma_1), \tag{2.11}$$

$$\boldsymbol{Z}_{2,j} \,|\, \gamma_2 \sim \mathrm{CRP}(\gamma_2), \tag{2.12}$$

$$\eta_{k,l} \,|\, \beta \sim \mathrm{Beta}(\beta, \beta), \tag{2.13}$$

$$R_{i,j} \,|\, \boldsymbol{Z}_{1,i}, \boldsymbol{Z}_{2,j}, \boldsymbol{\eta} \sim \mathrm{Bernoulli}(\boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j}), \tag{2.14}$$

where γ_1 and γ_2 are the concentration parameters for CRPs. Figure 2.4 shows the graphical representation of the IRM. As we can see, the difference between the generative processes of the SBM and IRM lies in the prior distributions for cluster assignment. In the CRP, given $\boldsymbol{z}_{1,-i}$, the posterior probability $P(z_{1,i} \,|\, \boldsymbol{z}_{1,-i})$ that i-th row object is assigned to cluster k^* is given as follows:

$$P(z_{1,i} = k^* \,|\, \boldsymbol{z}_{1,-i}) \propto \begin{cases} m_{1,-i,k^*}, & (\text{if } m_{1,-i,k^*} > 0) \\ \gamma_1. & (\text{if } k^* \text{ is a new cluster}) \end{cases} \tag{2.15}$$

As Eq. (2.15) shows, the assignment $z_{1,i}$ basically depends on the probability in proportion to the number of objects that is assigned to each cluster. However, a new cluster is generated by a probability in proportion to γ_1. Therefore, in the IRM, the numbers of clusters K and L change stochastically in each iteration. In summary, the posterior

for updating $z_{1,i}$ for the IRM is derived as follows:

$$P(z_{1,i} = k^* \mid \boldsymbol{z}_{1,-i}, \boldsymbol{z}_2, \boldsymbol{R}) \propto \begin{cases} m_{1,-i,k^*} \times \displaystyle\prod_{l=1}^{L} \frac{B(m^{+i}_{k^*,l} + \beta, \overline{m}^{+i}_{k^*,l} + \beta)}{B(m^{-i}_{k^*,l} + \beta, \overline{m}^{-i}_{k^*,l} + \beta)}, & (\text{if } m_{1,-i,k^*} > 0) \\ \gamma_1 \times \displaystyle\prod_{l=1}^{L} \frac{B(m^{+i}_{k^*,l} + \beta, \overline{m}^{+i}_{k^*,l} + \beta)}{B(\beta, \beta)}. & (\text{if } m_{1,-i,k^*} = 0) \end{cases} \tag{2.16}$$

Note that the EAP estimation of the link probability $\eta^{\text{EAP}}_{k,l}$ for the IRM can also be computed using Eq. (2.10).

2.4 Drawback of the Standard Biclustering Models

As described in chapter, in both the SBM and the IRM, a link probability between object i and object j is conditioned on $\boldsymbol{Z}^{\top}_{1,i}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$. That is, these models commonly assume that the link probability between two individual objects depends only on their cluster assignments $z_{1,i}$ and $z_{2,j}$. Many other extension of the SBM [72, 1, 43, 56] also follow this assumption.

However, the models with this assumption often obtain unexpected solutions. In other words, when analyzing real-world relational data, these IRM families often discover many small or sparse clusters. One of our major objectives in analyzing real-world relationships is to obtain insights into major segments and their interactions, meaning considering POS data, we want to know the following:

- major groups of customers with same preferences,
- major segments of items,
- pairs of a customer group and an item group which is strongly connected.

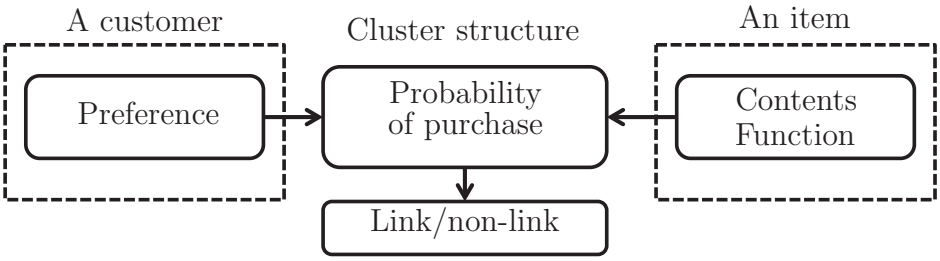
Therefore, both the small sized clusters and sparse clusters, which are not connected to any other primary clusters, are of no interest.

Why do the IRM and its families find unexpected solutions? As observed in the introduction, this is because real-world co-cluster structures are often destructed by *structured noise*. For example, the link probability related to a customer with a smaller budget decreases regardless of his or her preference compared with another customer with a larger budget. Similarly, a recently published book might be purchased by fewer customers regardless of its content compared with a long-standing book. In such cases, the related observations contain many non-links that must be regarded as noise. However such noise is distributed in a non-uniform manner that depends on related objects. Therefore, to obtain clear co-cluster structure, we must consider that observed relational data depend on both co-cluster structure and structured noise. Conventional models, which assume only co-cluster structure, can not distinguish between structured noise and co-cluster structure. Therefore, the IRM families find unexpected solutions.

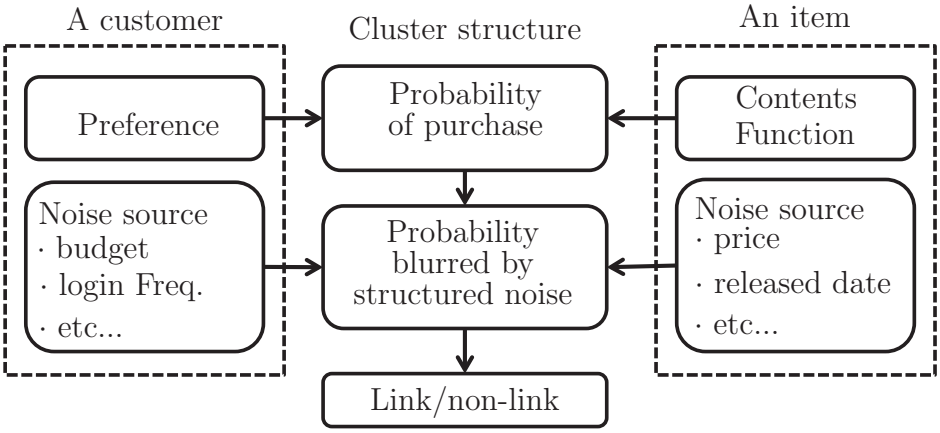
Chapter 3

Relevance Modeling with Logical Functions

In this chapter, we discuss the relevance-dependency modeling using Boolean functions. More specifically, for each entry of given relational data, we introduce a latent binary variable that indicates whether the entry relates to the block structure (foreground distribution) or to a background noise (background distribution). Then, we introduce a mechanism that the binary variable for an entry is determined by calculating an arbitrary Boolean function of Bernoulli trials from row and column objects. By introducing probability parameter that controls Bernoulli trial of corresponding object, the probability can be interpreted as the relevance parameter for the object. Thus, a probability for the Bernoulli trial can be interpreted as the relevance parameter for the object. By incorporating the mechanism, we propose an extension of the IRM termed the Relevance-Dependent IRM (RDIRM). The RDIRM is also an instance of Bayesian nonparametric models. Therefore, the RDIRM can automatically estimate the number of clusters. Furthermore, thanks to the conjugacy between component distributions, posterior inference for the RDIRM can be performed efficiently using collapsed Gibbs sampler. Experiments on real-world datasets show that our model extracts a clear



(a) SBMs' assumption



(b) Our assumption

Figure 3.1: Diagrams of (a) SBMs and (b) our assumptions.

bicluster structure. Moreover, we confirm that the estimated relevance values enable us to extract representative objects for each cluster.

3.1 Motivations

As we have discussed in Chapter 2, the SBM, IRM, and its families commonly assume that a link probability between two individual objects depends only on their cluster assignments (Fig. 3.1a). However, models with this assumption often output unexpected solutions. This is because, in real-world relationships, the underlying bicluster

structures are often destroyed by *structured noise* that blurs the underlying bicluster structure with individually different probabilities depending on the pair of related objects. For example, compared with customers with larger budgets, link probabilities related to those with smaller budgets decrease regardless of their preferences. As another example, link probabilities related to recently published books, in comparison with long-standing books, might reflect purchases by smaller numbers of customers regardless of their contents or topics. In these cases, the related observations contain many non-links which should be regarded as noise. Unfortunately, the amount of such noise depends on the related objects. In other words, this noise has a stochastic structure that destroys the underlying bicluster structure in a non-uniform manner (Fig. 3.1b). The conventional models, which assume only bicluster structure, can not distinguish between the structured noise and bicluster structure. Therefore, to obtain informative bicluster structures that clearly indicate each customer's preference and each item's functions, we need to consider both the bicluster structure and structured noise underlying real-world relational data.

To overcome the structured noise problem, we propose a new probabilistic mechanism called the *relevance dependent mechanism* that describes the generative process for the structured noise. In our mechanism, a relevance parameter ($\in [0, 1]$) is introduced for each object. The relevance parameter controls how closely the object's links follow the bicluster structure. That is, links related to an object with high relevance (close to 1.0) primarily follow the bicluster structure. Conversely, links related to an object with low relevance (close to 0.0) primarily originate from the noise source. For example, customers with larger budgets can purchase whatever they needs. As another example, long standing books might be acknowledged by many customers. Therefore, their related observations (links and non-links) might be relevant to a customer's preference or the topic of a book. In such cases, the relevance parameters for them become close to 1.0. As a result, their related observations are primarily explained by bicluster

structure. On the other hand, considering customers with smaller budgets, the related observations might contain many non-links because such customers can purchase only a few items even if they need many items. Similarly, recently published books might be acknowledged by a small number of customers. Therefore, observations related to the books might also contain many non-links. In these cases, the relevance parameters for them become close to 0.0. Consequently, these many non-links are explained by the noise source in our mechanism. Since our mechanism strictly follows the principles of hierarchical Bayesian modeling, we can incorporate the mechanism into arbitrary conventional Bayesian biclustering models. Moreover, we propose a new biclustering model, the *Relevance Dependent Infinite Relational Model* (RDIRM), incorporating our new mechanism into the IRM. In the RDIRM, observed relational data is factorized into a bicluster structure and structured noise. Consequently, we can obtain not only a clear bicluster structure but also a relevance level for each object that reflects its relevance to the bicluster structure.

3.2 Relevance Dependent Infinite Relational Model (RDIRM)

In this section, we propose relevance dependent mechanism that gives a generative process for modeling structured noise in relational data. We also propose a new biclustering model, called the *Relevance Dependent Infinite Relational Model* (RDIRM), and review some related works.

3.2.1 The Relevance Dependent Mechanism

To describe the generative process in which underlying bicluster structure is blurred by a noise source, we consider a background link probability η_0. We assume that

each observation within relational data is generated from a mixture distribution of cluster dependent link probability $\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$ and the background probability η_0. Then, to describe the situation that noise level depends on the corresponding objects, we introduce *relevance parameters* $\rho_{1,i}, \rho_{2,j} \in [0, 1]$ for each object and construct relevance dependent mechanism as follows:

$$r_{1,i\to j} \,|\, \rho_{1,i} \sim \text{Bernoulli}(\rho_{1,i}), \tag{3.1}$$

$$r_{2,j\to i} \,|\, \rho_{2,j} \sim \text{Bernoulli}(\rho_{2,j}), \tag{3.2}$$

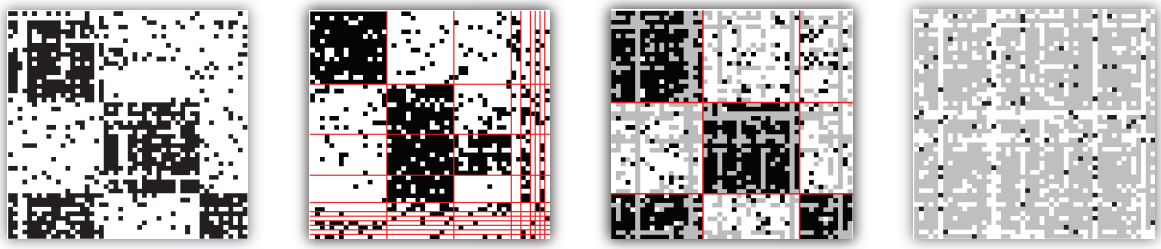
$$r_{i,j} = f(r_{1,i\to j}, r_{2,j\to i}), \tag{3.3}$$

$$\xi_{i,j} = r_{i,j} \times \boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j} + (1 - r_{i,j}) \times \eta_0, \tag{3.4}$$

where $f(\cdot, \cdot)$ is an arbitrary Boolean function that returns one or zero. In this mechanism, the link probability for an entry (i, j) is described either the foreground probability $\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$ or the background probability η_0 depending on their relevance parameters $\rho_{1,i}, \rho_{2,j}$. For example, when f is a logical sum, it corresponds to assuming the mixture rate to be $1-(1-\rho_{1,i})(1-\rho_{2,j})$. When f is a logical product, the mixture rate becomes $\rho_{1,i}\times\rho_{2,j}$. Using the proposed mechanism, we can describe the situation that underlying bicluster structure is blurred by structured noise. That is, if $\rho_{1,i}$ and $\rho_{2,j}$ become close to 1.0, the corresponding observation $R_{i,j}$ follows to the cluster structure. Conversely, if $\rho_{1,i}$ and $\rho_{2,j}$ decrease to 0.0, the link probability is blurred by background probability η_0. Therefore, by incorporating this relevance dependent mechanism into conventional models, we can construct a new biclustering model that can factorize relational data into a clear bicluster structure and structured noise.

3.2.2 The Relevance Dependent Infinite Relational Model

Here, we propose a new biclustering model called the RDIRM, which is an extension of the IRM incorporating the relevance dependent mechanism we described in 3.2.1. In the RDIRM, a link probability $\xi_{i,j}$ between object i and object j follows the relevance



(a) Toy data (b) IRM (c) RDIRM:fg (d) RDIRM:bg

Figure 3.2: Toy data example: (a) synthetic 50×50 relational data (white corresponds to zero, and black corresponds to one); (b) IRM solution (rows and columns are sorted by obtained cluster indices); (c) and (d) RDIRM solutions: (c) shows the area assigned to foreground ($r_{i,j} = 1$), and (d) shows the area assigned to background ($r_{i,j} = 0$) (gray area indicates that the corresponding entries are assigned to another layer).

dependent mechanism, and the foreground link probability $\boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j}$ follows the IRM. The full description of the generative model for the RDIRM is as follows:

$$\boldsymbol{Z}_{1,i} \,|\, \gamma_1 \sim \mathrm{CRP}(\gamma_1), \;\; \boldsymbol{Z}_{2,j} \,|\, \gamma_2 \sim \mathrm{CRP}(\gamma_2), \tag{3.5}$$

$$\eta_{k,l} \,|\, \beta \sim \mathrm{Beta}(\beta, \beta), \;\; \eta_0 \,|\, \beta \sim \mathrm{Beta}(\beta_0, \beta_0), \tag{3.6}$$

$$\rho_{1,i} \,|\, \beta_1 \sim \mathrm{Beta}(\beta_1, \beta_1), \tag{3.7}$$

$$\rho_{2,j} \,|\, \beta_2 \sim \mathrm{Beta}(\beta_2, \beta_2), \tag{3.8}$$

$$r_{1,i \to j} \,|\, \rho_{1,i} \sim \mathrm{Bernoulli}(\rho_{1,i}), \tag{3.9}$$

$$r_{2,j \to i} \,|\, \rho_{2,j} \sim \mathrm{Bernoulli}(\rho_{2,j}), \tag{3.10}$$

$$r_{i,j} = f(r_{1,i \to j}, r_{2,j \to i}), \tag{3.11}$$

$$\xi_{i,j} = r_{i,j} \times \boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j} + (1 - r_{i,j}) \times \eta_0, \tag{3.12}$$

$$R_{i,j} \,|\, \xi_{i,j} \sim \mathrm{Bernoulli}(\xi_{i,j}). \tag{3.13}$$

Figure 3.3 graphically represents this model.

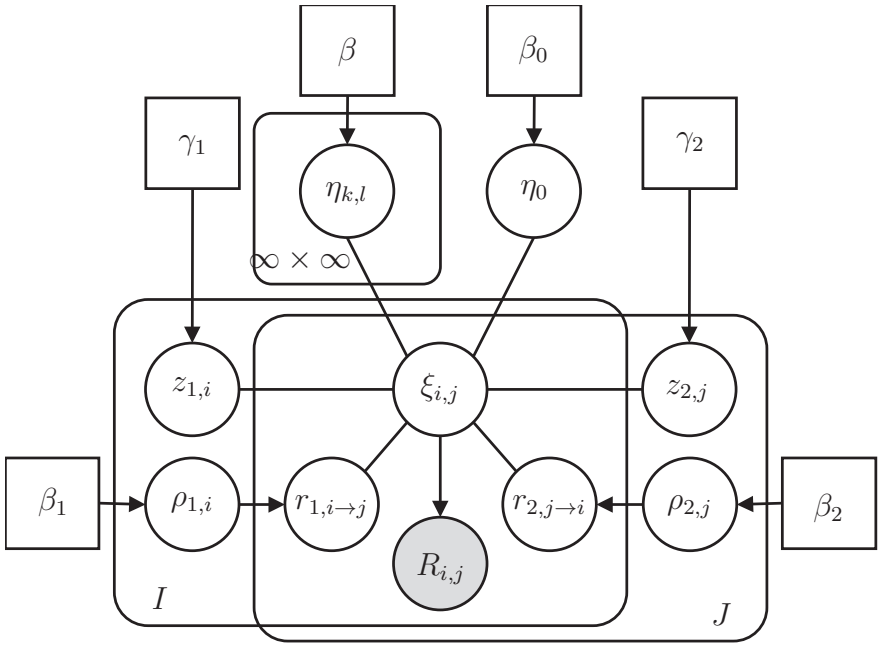


Figure 3.3: Graphical representations for the generative models of the RDIRM. Circle nodes denote random variables, square nodes denote hyperparameters, shaded nodes denote observations, and round-edged squares indicate number of individual variables. Directed connections denote probabilistic dependencies, and non-directed connections denote dependencies determined by an arithmetic function.

Now, we explain briefly the generative process of the RDIRM. First, the cluster assignments $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are given as in the original IRM (Eq. (3.5)). Second, the foreground distribution $\eta_{k,l}$ and the background distribution η_0 are independently drawn from corresponding beta priors (Eq. (3.6)). Third, for each object, the relevance parameters $\rho_{1,i}$ and $\rho_{2,j}$ are given from beta priors (Eqs. (3.7) and (3.8)). Fourth, the two binary variables $r_{1,i\to j}$ and $r_{2,j\to i}$ are drawn by Bernoulli trials with corresponding relevance values (Eqs. (3.9) and (3.10)). Fifth, either the foreground $\eta_{k,l}$ or the background η_0 is selected by the interaction of $r_{1,i\to j}$ and $r_{2,j\to i}$ via logical function f (Eqs. (3.11) and (3.12)). Finally, an entry $R_{i,j}$ is generated from the selected probability

(Eq. (3.13)).

The difference between our RDIRM and the original IRM is that our RDIRM describes a generative process for relationships with structured noise by introducing objects' relevance parameters and their interaction mechanism. That is, our RDIRM can bicluster relational data based on a subset of observations that comes from the underlying bicluster structure.

When f is a logical sum, an observation is drawn from the foreground when at least one of the related objects $O_{1,i}$ or $O_{2,j}$ has high relevance. This models situations in which entries from the bicluster structure can be generated by a one-sided request, such as sending an e-mail or following a hyperlink on the Internet. In contrast, when f is a logical product, the entry follows the bicluster structure only when the related objects cooperate with each other. This models situations in which an object i that aims to have a link with another object j can be constrained by the relevance of object j. Certainly, we can adopt other logical functions for other interaction models.

For an intuitive understanding of our RDIRM, we show a toy example. Figure 3.2a shows hand-constructed relational data. As we can see in Fig. 3.2a, there is a 3×3 bicluster structure in the data. However, we can also see that the cluster structure is blurred by structured noise. Figures 3.2b-3.2d show the solutions obtained by the IRM and the RDIRM with $f(\cdot, \cdot)$ as the logical product. As Fig. 3.2b shows, the IRM fails to extract true partitions, because it assumes that all observations are relevant to the underlying cluster structure. In contrast, the RDIRM (Figs. 3.2c and 3.2d) finds true partitions by considering the mixture of the cluster structure and the structured noise. Note that the inference algorithm for the RDIRM is detailed in Section 3.3.

3.2.3 Related Works

First, we briefly review some related work. Next, we show that our RDIRM can be viewed as a generalization of some conventional models.

Many studies of statistical models for extracting hidden low-dimensional representations from matrix-represented relational data exist [59, 60, 7, 47, 34]. These models are useful for predicting missing entries, an important task in many application domains such as recommendation systems or collaborative filtering. Among these models, matrix factorization based approaches are known to have the better predicting ability [59, 60, 66].

Another motivation for modeling relational data is to find an interpretable structure underlying the data. The Latent Dirichlet Allocation (LDA) [7] is based on K-mixture models, which assume a mixture of K component distributions for each object within data. In general, K-mixture models assume a Dirichlet prior, defined on a K-dimensional simplex, for the objects' latent variables. Therefore, using a mixture model based approach, we can obtain more interpretable insights from data, greatly facilitating exploratory data analysis.

The SBM and IRM are extended K-mixture models for relational data. These models assume K and L mixture components for row objects and column objects, respectively. Therefore, these models are suitable for analyzing directed networks (e.g., e-mail transactions) and multi-domain relationships (e.g., relationships between users and items). Especially, the IRM has the great advantage that it can automatically estimate the number of mixture components K and L (so called ∞-mixture model). At present, our major motivation is to extract an interpretable cluster structure from blurred relational data; thus, we focus on the ∞-mixture based model in this paper.

Recently, there have been many studies aimed at improving the performance of the above mentioned models. To improve predicting accuracy, [64, 40, 18, 42, 72] extended the conventional models to ensure the models utilize side information such as meta information or partially observed supervisory variables.

Another challenge in this field is to handle noisy data. In real-world data, an ideal cluster structure is often blurred by a large number of irrelevant entries. Unfortunately,

such noisy entries are often distributed in a non-uniform manner. That is, some objects clearly follow a cluster structure, while other objects contain large number of irrelevant entries. In such cases, not only relevant entries but also irrelevant entries follow some probabilistic structure. Therefore, it is important to consider a generative mechanism to extract a clear cluster structure from blurred real-world relationships.

Our RDIRM can be viewed as one of the noise filtering models called *subset models* that assume only a part of the observations are relevant to underlying cluster structure. These subset models commonly assume a background probability to describe irrelevant entries. Clustering models that consider the influence of irrelevant entries were first discussed by Newton [46] and Hoff [25] relative to clustering biological sequences.

For biclustering relational data, there have been an extensions of the IRM in which the background probability affects link probabilities. The Subset IRM (SIRM) proposed by Ishiguro *et al.* [33] also considers a generative model in which link probability is a mixture distribution of a foreground probability $\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$ and a background probability η_0. In the SIRM, binary variables $s_{1,i}, s_{2,j} \in \{0,1\}$ are introduced to indicate whether each object is relevant to the underlying bicluster structure. Then, a subset of $\boldsymbol{R}$, where $s_{1,i} \times s_{2,j} = 1$, is explained by the clustering model $\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$, while the rest are explained by the background probability η_0. In the SIRM, specifically, the link $R_{i,j}$ is drawn as follows:

$$
\begin{aligned}
&R_{i,j} \mid \boldsymbol{Z}_{1,i}, \boldsymbol{Z}_{2,j}, s_{1,i}, s_{2,j}, \boldsymbol{\eta}, \eta_0 \\
&\quad \sim \text{Bernoulli}\left(\begin{pmatrix} \mathbb{I}(s_{1,i} \times s_{2,j} = 1) \\ \mathbb{I}(s_{1,i} \times s_{2,j} = 0) \end{pmatrix}^{\top} \begin{pmatrix} \boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j} \\ \eta_0 \end{pmatrix}\right),
\end{aligned} \tag{3.14}
$$

where $\mathbb{I}(\cdot)$ is one if the predicate holds and zero otherwise.

The proposed RDIRM can be viewed as a generalization of the SIRM. In the RDIRM, continuous relevance parameters $\rho_{1,i}, \rho_{2,j} \in [0,1]$ are introduced instead of the $s_{1,i}$ and $s_{2,j}$ in the SIRM. Consequently, the RDIRM can estimate the confidence

that an object is relevant to the underlying cluster structure. To clarify the relationships between our RDIRM and the SIRM, the link probability of the RDIRM can be rewritten equivalently as follows:

$$
\begin{aligned}
r_{1,i\to j} \mid \rho_{1,i} \sim & \text{Bernoulli}(\rho_{1,i}), \\
r_{2,j\to i} \mid \rho_{2,j} \sim & \text{Bernoulli}(\rho_{2,j}), \\
R_{i,j} \mid \boldsymbol{Z}_{1,i}, \boldsymbol{Z}_{2,j}, r_{1,i\to j}, r_{2,j\to i}, \boldsymbol{\eta}, \eta_0 & \\
\sim & \text{Bernoulli}\left(\begin{pmatrix} f(r_{1,i\to j}, r_{2,j\to i}) \\ 1 - f(r_{1,i\to j}, r_{2,j\to i}) \end{pmatrix}^{\top} \begin{pmatrix} \boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j} \\ \eta_0 \end{pmatrix}\right). \qquad (3.15)
\end{aligned}
$$

Note that these equations are equivalent to the SIRM when the relevance parameters are constrained to have $\rho_{1,i}, \rho_{2,j} \in \{0.0, 1.0\}$ and the Boolean function f is the logical product. As is evident from the forms of Eqs. (3.14) and (3.15), we can see that our RDIRM is a natural generalization of the SIRM.

3.3 Inference

We also use the Collapsed Gibbs Sampler to infer the parameters of the RDIRM. Given $r_{i,j}$, the relational data $\boldsymbol{R}$ are separated into foreground and background parts; thus, the relevance parameters $\rho_{1,i}, \rho_{2,j}$ and the link probabilities $\eta_{k,l}, \eta_0$ can be integrated out. Therefore, the inference for the RDIRM is performed by sampling the assignments $\boldsymbol{z}_1, \boldsymbol{z}_2$ and the binary variables $\boldsymbol{r}_1, \boldsymbol{r}_2$ one after the other. In this section, we show the derived posteriors for running Gibbs sampling.

3.3.1 Sampling Cluster Assignments $\boldsymbol{z}_1, \boldsymbol{z}_2$

Because $\boldsymbol{z}_2$ can be sampled in the same manner as $\boldsymbol{z}_1$, we concentrate on $\boldsymbol{z}_1$. We can assume that the switch variables $\boldsymbol{r}$ ($\boldsymbol{r}_1$ and $\boldsymbol{r}_2$) have already been given before taking a

sample of $z_{1,i}$. Given $\boldsymbol{r}$, the cluster assignments depend only on the foreground part of the observations. Therefore, the conditional posterior for $z_{1,i} = k^*$ is derived as follows:

$$P(z_{1,i} = k^* \mid \boldsymbol{z}_{1,-i}, \boldsymbol{z}_2, \boldsymbol{r}, \boldsymbol{R}) \propto \begin{cases} m_{1,-i,k^*} \times \displaystyle\prod_{l=1}^{L} \frac{B(m^{+i}_{k^*,l:f} + \beta, \overline{m}^{+i}_{k^*,l:f} + \beta)}{B(m^{-i}_{k^*,l:f} + \beta, \overline{m}^{-i}_{k^*,l} + \beta)}, & (\text{if } m_{1,-i,k^*} > 0) \\ \gamma_1 \times \displaystyle\prod_{l=1}^{L} \frac{B(m^{+i}_{k^*,l:f} + \beta, \overline{m}^{+i}_{k^*,l:f} + \beta)}{B(\beta, \beta)}, & (\text{if } m_{1,-i,k^*} = 0) \end{cases} \tag{3.16}$$

where the counts $m_{\cdot,\cdot:f}$ and $\overline{m}_{\cdot,\cdot:f}$ denote the number of links and non-links for which $r_{i,j} = 1$, and are computed as follows:

$$m^{+i}_{k^*,l:f} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^*(z_{1,i} := k^*), z_{2,j} = l}} \sum R_{x,j} \times r_{x,j}, \quad \overline{m}^{+i}_{k^*,l:f} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^*(z_{1,i} := k^*), z_{2,j} = l}} \sum \overline{R}_{x,j} \times r_{x,j},$$

$$m^{-i}_{k^*,l:f} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^*(x \neq i), z_{2,j} = l}} \sum R_{x,j} \times r_{x,j}, \quad \overline{m}^{-i}_{k^*,l:f} = \sum_{\substack{x \in T_1, j \in T_2: \\ z_{1,x} = k^*(x \neq i), z_{2,j} = l}} \sum \overline{R}_{x,j} \times r_{x,j}.$$

Note that if $r_{i,j} = 1$ for all (i, j),then Eq. (3.16) is equivalent to the original IRM's sampler.

3.3.2 Sampling Switch Variables $r_{1,i \to j}, r_{2,j \to i}$

As the sampling of $r_{2,j \to i}$ can be performed in the same manner as the sampling of $r_{1,i \to j}$, we concentrate on $r_{1,i \to j}$. Given $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, we have a finite number $K \times L$ of clusters. Thus, the conditional posterior for $r_{1,i \to j}$ is derived as follows:

$$\begin{aligned} &P(r_{1,i \to j} \mid \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}_{1,-(i \to j)}, \boldsymbol{r}_2, \boldsymbol{R}) \\ &\quad \propto P(R_{i,j} \mid r_{1,i \to j}, \boldsymbol{r}_{1,-(i \to j)}, \boldsymbol{r}_2, \boldsymbol{R}_{-(i,j)})^{1 - f(r_{1,i \to j}, r_{2,j \to i})} \\ &\qquad \times P(R_{i,j} \mid \boldsymbol{z}_1, \boldsymbol{z}_2, r_{1,i \to j}, \boldsymbol{r}_{1,-(i \to j)}, \boldsymbol{r}_2, \boldsymbol{R}_{-(i,j)})^{f(r_{1,i \to j}, r_{1,j \to i})} \\ &\qquad \times P(r_{1,i \to j} \mid \boldsymbol{r}_{1,i \to (-j)}), \end{aligned} \tag{3.17}$$

where $\boldsymbol{R}_{-(i,j)}$ denotes the entire set of $\boldsymbol{R}$ excluding $R_{i,j}$. Similarly, $\boldsymbol{r}_{1,-(i \to j)}$ denotes the entire set of $\boldsymbol{r}_1$ without $r_{1,i \to j}$, and $\boldsymbol{r}_{1,i \to (-j)}$ denotes a vector of $r_{1,i \to t}$s that are

related to object i without $r_{1,i\to j}$. The terms on the right-hand side of Eq. (3.17) are computed as follows:

$$P(R_{i,j} \mid r_{1,i\to j}, \boldsymbol{r}_{1,-(i\to j)}, \boldsymbol{r}_2, \boldsymbol{R}_{-(i,j)}) = \frac{(m_{\overline{f}}^{-(i,j)} + \beta_0)^{R_{i,j}} (\overline{m}_{\overline{f}}^{-(i,j)} + \beta_0)^{1-R_{i,j}}}{m_{\overline{f}}^{-(i,j)} + \overline{m}_{\overline{f}}^{-(i,j)} + 2\beta_0}, \tag{3.18}$$

$$P(R_{i,j} \mid \boldsymbol{z}_1, \boldsymbol{z}_2, r_{1,i\to j}, \boldsymbol{r}_{1,-(i\to j)}, \boldsymbol{r}_2, \boldsymbol{R}_{-(i,j)}) = \frac{(m_{k,l:f}^{-(i,j)} + \beta)^{R_{i,j}} (\overline{m}_{k,l:f}^{-(i,j)} + \beta)^{1-R_{i,j}}}{m_{k,l:f}^{-(i,j)} + \overline{m}_{k,l:f}^{-(i,j)} + 2\beta}, \tag{3.19}$$

$$P(r_{1,i\to j} \mid \boldsymbol{r}_{1,i\to(-j)}) = \frac{(n_{\boldsymbol{r}_{1,i}}^{-(i,j)} + \beta_1)^{r_{1,i\to j}} (n_{\overline{\boldsymbol{r}}_{1,i}}^{-(i,j)} + \beta_1)^{1-r_{1,i\to j}}}{J - 1 + 2\beta_1}, \tag{3.20}$$

where $m_{\overline{f}}^{-(i,j)}$ and $\overline{m}_{\overline{f}}^{-(i,j)}$ denote the numbers of links and non-links, respectively, such that $r_{s,t} = 0$ for all pairs $(s,t) \neq (i,j)$; $m_{k,l:f}^{-(i,j)}$ and $\overline{m}_{k,l:f}^{-(i,j)}$ denote the numbers of links and non-links, respectively, such that $z_{1,s} = k$, $z_{2,t} = l$ and $r_{s,t} = 1$ for all pairs $(s,t) \neq (i,j)$; and $n_{\boldsymbol{r}_{1,i}}^{-(i,j)}$ and $n_{\overline{\boldsymbol{r}}_{1,i}}^{-(i,j)}$ denote the numbers of $r_{1,i\to t} = 1\{t \neq j\}$ and $r_{1,i\to t} = 0\{t \neq j\}$, respectively, within $\boldsymbol{r}_{1,i\to(-j)}$. Specifically, these counts are computed as follows:

$$\begin{aligned}
n_{\boldsymbol{r}_{1,i}}^{-(i,j)} &= \sum_{t\in T_2: t\neq j} r_{1,i\to t}, \quad n_{\overline{\boldsymbol{r}}_{1,i}}^{-(i,j)} = \sum_{t\in T_2: t\neq j} (1 - r_{1,i\to t}), \\
m_{\overline{f}}^{-(i,j)} &= \sum_{\substack{s\in T_1, t\in T_2: \\ (s,t)\neq(i,j)}} \sum R_{s,t} \times (1 - f(r_{1,s\to t}, r_{2,t\to s})), \\
\overline{m}_{\overline{f}}^{-(i,j)} &= \sum_{\substack{s\in T_1, t\in T_2: \\ (s,t)\neq(i,j)}} \sum (1 - R_{s,t}) \times (1 - f(r_{1,s\to t}, r_{2,t\to s})), \\
m_{k,l:f}^{-(i,j)} &= \sum_{\substack{s\in T_1, t\in T_2: \\ z_{1,s}=k, z_{2,t}=l, (s,t)\neq(i,j)}} \sum R_{s,t} \times f(r_{1,s\to t}, r_{2,t\to s}), \\
\overline{m}_{k,l:f}^{-(i,j)} &= \sum_{\substack{s\in T_1, t\in T_2: \\ z_{1,s}=k, z_{2,t}=l, (s,t)\neq(i,j)}} \sum (1 - R_{s,t}) \times f(r_{1,s\to t}, r_{2,t\to s}),
\end{aligned}$$

respectively.

3.3.3 Estimations for $\eta_{k,l}$, η_0, $\rho_{1,i}$,and $\rho_{2,j}$

The EAP estimations for the marginalized parameters $\eta_{k,l}$, η_0, $\rho_{1,i}$, and $\rho_{2,j}$ are computed as follows:

$$\eta_{k,l}^{\text{EAP}} = \frac{m_{k,l:f} + \beta}{m_{k,l:f} + \overline{m}_{k,l,f} + 2\beta}, \tag{3.21}$$

$$\eta_0^{\text{EAP}} = \frac{m_{\overline{f}} + \beta_0}{m_{\overline{f}} + \overline{m}_{\overline{f}} + 2\beta_0}, \tag{3.22}$$

$$\rho_{1,i}{}^{\text{EAP}} = \frac{n_{\boldsymbol{r}_{1,i}} + \beta_1}{J + 2\beta_1}, \quad \rho_{2,j}{}^{\text{EAP}} = \frac{n_{\boldsymbol{r}_{2,j}} + \beta_2}{I + 2\beta_2}, \tag{3.23}$$

where

$$m_{k,l:f} = \sum_{\substack{i\in T_1, j\in T_2: \\ z_{1,i}=k, z_{2,t}=l}} \sum R_{i,j} \times f(r_{1,i\to j}, r_{2,j\to i}),$$

$$\overline{m}_{k,l:f} = \sum_{\substack{i\in T_1, j\in T_2: \\ z_{1,i}=k, z_{2,t}=l}} \sum (1 - R_{i,j}) \times f(r_{1,i\to j}, r_{2,j\to i}),$$

$$m_{\overline{f}} = \sum_{i\in T_1, j\in T_2} \sum R_{i,j} \times (1 - f(r_{1,i\to j}, r_{2,j\to i})),$$

$$\overline{m}_{\overline{f}} = \sum_{i\in T_1, j\in T_2} \sum (1 - R_{i,j}) \times (1 - f(r_{1,i\to j}, r_{2,j\to i})),$$

$$n_{\boldsymbol{r}_{1,i}} = \sum_{j\in T_2} r_{1,i\to j}, \; n_{\boldsymbol{r}_{2,j}} = \sum_{i\in T_1} r_{2,j\to i}.$$

3.3.4 Selecting Boolean Function f

In our relevance dependent mechanism, the problem of selecting Boolean function $f(\cdot,\cdot)$ remains. In analyzing real-world data, we might often encounter the situation that we have no prior knowledge available for selecting a Boolean function. Therefore, it is desirable to simultaneously estimate the form of the Boolean function and the model parameters. Although we postpone the essential study of this problem for our future work, we discuss some heuristic strategies here.

One solution is to compare the values of the log likelihood of trained models for several Boolean functions. A higher value of log likelihood indicates that the model fits

the training dataset better; thus, it might be a reasonable criterion. Another approach is to compare the sizes of estimated cluster blocks $K \times L$. If a Boolean function successfully estimates the noise source for a given dataset, the RDIRM might extract a simple and clear bicluster structure. Therefore, selecting a Boolean function by the number of estimated cluster blocks might also be a reasonable strategy.

Although these approaches work to some extent, they sometimes give different results. Therefore, at this time, we select the Boolean function according to the background knowledge for datasets.

3.4 Experiments

In this section, we present our experimental results. Although the comparison between the SIRM and our RDIRM might be interesting, we compare the performance of the RDIRM with that of the original IRM to clarify the effectiveness of the relevance dependent mechanism. [1] Through all the experiments, we assume that the priors of all binary variables in the generative models are uniform (Beta$(1.0, 1.0)$). In addition, we estimate the concentration parameters γ_1, γ_2 for the DPs assuming gamma priors by sampling method presented in [16].

3.4.1 Experiments on Synthetic Datasets

We prepared 12 synthetic datasets. First, according to the generative model of the RDIRM, we created five synthetic datasets, Data1(0.0), Data1(0.2), Data1(0.5), Data1(0.8), and Data1(1.0), where the numbers in parentheses indicate the background link prob-

[1] In the SIRM, a part of objects are removed for the clustering targets. However, considering a situation where in the clustering result is used for recommendation, every object must be assigned to its nearest cluster. Hence, the SIRM does not meet our requirement. Therefore, we compare the RDIRM with the original IRM.

abilities η_0 for the datasets. We set the logical function f for the RDIRM to be the logical sum. The cluster assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ were independently generated from fixed-dimensional categorical distributions. The parameter values used for generating the datasets were $I = J = 200$, $\beta = (0.5, 0.5)$, and $\beta_1 = \beta_2 = (4.0, 3.0)$; the number of clusters were set as $K = 4$ and $L = 5$; and the parameters for the categorical distributions were $(0.4, 0.3, 0.2, 0.1)$ and $(0.33, 0.27, 0.20, 0.13, 0.07)$ for T_1 and T_2, respectively. Next, we also created five synthetic datasets in a similar manner (from Data2(0.0) to Data2(1.0)), except that we set the logical function f to be the logical product and set both β_1 and β_2 to be $(4.0, 2.0)$. Finally, we created two datasets without background influences, (Data1(NULL) and Data2(NULL)). We applied the logical sum version of the RDIRM to Data1 and the logical product version to Data2, respectively.

We use three measurements to evaluate clustering performance. One is the Adjusted Rand Index (ARI) [29], which is widely used for computing the similarity between true and estimated clustering results. The ARI takes a value in the range 0.0 – 1.0, taking the value 1.0 when a clustering result is exactly equivalent to ground truth. The second is the number of erroneous estimated clusters (EC). We computed the average of these measures for the two sets T_1 and T_2. The third is the test data log likelihood (TDLL), which indicates the predictive robustness of a generative model. We hid 1.0% of the observations during inference (keeping it small so that the latent cluster structure do not change), and measured the averaged log likelihood that a hidden entry takes the actual value. A larger value is better, and a smaller one means that the model overfits the data. Finally, we repeated the experiment 10 times for each dataset using different random seeds to find an overall average.

Table 3.1 lists the computed measures. For every dataset except Data1(NULL) and Data2(NULL), we confirm that the RDIRM outperformed the IRM. In particular, the RDIRM maintained good performance for sparse ($\eta_0 \approx 0.0$) or dense ($\eta_0 \approx 1.0$) data. We also list in Table 3.2 the EAP estimations of the background probability η_0^{EAP} and

Table 3.1: ARI, EC, and TDLL on synthetic datasets.

	ARI		EC		TDLL	
Dataset	IRM	RDIRM	IRM	RDIRM	IRM	RDIRM
Data1(NULL)	**1.000**	0.999	**0.000**	0.030	-0.302	**-0.261**
Data1(0.0)	0.712	**0.999**	0.678	**0.022**	-0.410	**-0.315**
Data1(0.2)	0.806	**1.000**	0.480	**0.010**	-0.432	**-0.363**
Data1(0.5)	0.868	**0.993**	0.270	**0.090**	-0.459	**-0.405**
Data1(0.8)	0.834	**0.999**	0.388	**0.013**	-0.462	**-0.385**
Data1(1.0)	0.806	**0.999**	0.435	**0.025**	-0.425	**-0.330**
Data2(NULL)	**1.000**	0.996	**0.000**	**0.000**	-0.316	**-0.232**
Data2(0.0)	0.629	**0.980**	1.053	**0.020**	-0.424	**-0.196**
Data2(0.2)	0.627	**0.913**	0.735	**0.105**	-0.576	**-0.431**
Data2(0.5)	0.759	**0.930**	0.488	**0.105**	-0.614	**-0.526**
Data2(0.8)	0.724	**0.917**	0.738	**0.097**	-0.558	**-0.438**
Data2(1.0)	0.644	**0.981**	0.910	**0.083**	-0.390	**-0.183**

the estimated ratios of the foreground for synthetic datasets except Data1(NULL) and Data2(NULL). The ground truths of the foreground ratios (FRs) are 0.8197 and 0.4622 for Data1 and Data2, respectively. As the table shows, the RDIRM performs well in estimating ground truths.

3.4.2 Experiments on Real-world Datasets

We applied the RDIRM to three real-world datasets. The first dataset is "MovieLens[2]", which contains a large number of user ratings of movies on a five-point scale. In our experiment, we created a binary relational dataset with a threshold that yields $R_{i,j} = 1$

[2]http://www.grouplens.org/, as of 2003.

Table 3.2: Estimated background probabilities (η_0^{EAP}) and the FRs.

Dataset	η_0^{EAP}	FR
Data1(0.0)	0.0085	0.8484
Data1(0.2)	0.1970	0.8462
Data1(0.5)	0.4531	0.8588
Data1(0.8)	0.7674	0.8607
Data1(1.0)	0.9876	0.8611
Data2(0.0)	0.0022	0.4884
Data2(0.2)	0.2139	0.4548
Data2(0.5)	0.5033	0.4658
Data2(0.8)	0.7845	0.4397
Data2(1.0)	0.9872	0.4654

for ratings higher than 3 points and $R_{i,j} = 0$ for all other ratings. That is, an entry $R_{i,j} = 1$ indicates that user i likes movie j. There are a total of 943 users and 1,682 movies in the dataset, and 3.5% of the observations are links. The second dataset is "animal-feature" [55], which includes relationships between 50 mammals and 85 features. Each feature is rated on a scale of 0–100 for each animal. We prepared binary relational data with a threshold that yields $R_{i,j} = 1$ for all ratings higher than the average of the entire set of ratings (20.79). That is, we used the relational value $R_{i,j} = 1$ ($R_{i,j} = 0$) to indicate that animal i has (does not have) feature j. In this dataset, 36.8% of the relations are links. The last dataset is "Enron" [35], which contains e-mail transactions among the employees of the Enron company. We extracted e-mail transactions from October 2001, when the Enron accounting scandal was first reported. This dataset contains 149 Enron employees. For this dataset, $R_{i,j} = 1$ ($R_{i,j} = 0$) was used to indicate if an e-mail was (not) sent from employee i to employee j. In this dataset, 2.6% of the observations are links.

We used a logical sum version of the RDIRM for MovieLens and Enron. For animal-feature, we used a logical product version of the RDIRM. The reason for using the logical sum for MovieLens was that a user can watch any movie according to his or her preference, and movies are usually promoted independently of users. Similarly, for Enron, an employee can send e-mails to anyone if he/she would like to and one can receive e-mails from anyone who would like to send one. Therefore, it seemed natural that the foregrounds (relevant entries) for MovieLens and Enron are generated according to the user's (sender's) relevance $\rho_{1,i}$ or the movie's (receiver's) relevance $\rho_{2,j}$. In contrast, animals' features are acquired through evolutionary history depending on the type of animal. For example, aquatic features such as "swims" or "water" cannot be acquired by terrestrial animals. Therefore, the type of animal limits the features it can acquire, and similarly, the type of a feature limits the types of animals that can be related to that feature. Therefore, we used the logical product version of the RDIRM for the animal-feature dataset.

Figure 3.4 shows the clustering results and computed TDLLs for these real-world datasets. Figure 3.5 shows color maps for the estimated foreground probabilities $\eta_{k,l}^{\mathrm{EAP}}$. The background probabilities η_0 that the RDIRM estimated were 0.0000, 0.0036, and 0.0016 for the MovieLens, animal-feature, and Enron datasets, respectively. For MovieLens and animal-feature, we can see that the original IRM organized many non-informative cluster-blocks, because the IRM considered all entries to be relevant for the bicluster structure. In contrast, the RDIRM found more clear cluster structures by using the proposed relevance dependent mechanism, which selects an informative subset of entries via interactions of the objects' relevance values. For Enron, the number of obtained cluster blocks for the IRM and RDIRM are nearly equivalent. However, the RDIRM found more strongly connected cluster blocks (Figs. 3.5e and 3.5f). For all the datasets, the computed TDLLs show that the RDIRM outperformed the original IRM for both datasets in predicting hidden entries.

Additionally, for qualitative comparison, we list the examples of clusters obtained for MovieLens and animal-feature. [3] The left side of Table 3.3 lists examples of the movie clusters produced by the RDIRM for MovieLens. In the columns for the number of links and ${\rho_{2,j}}^{\mathrm{EAP}}$, we can see that ${\rho_{2,j}}^{\mathrm{EAP}}$ tends to increase with the number of links. This indicates that we can regard the relevance values as an indication of the popularity of the movies within the cluster. In contrast, the original IRM treats all links and non-links as relevant observations. Therefore, the cluster assignment for a movie is affected by not only the movie's category but also the number of related links. The right side of Table 3.3 lists examples of the feature clusters obtained by the RDIRM for the animal-feature dataset. As with the results for MovieLens, we can see that the estimated $\rho_{2,j}$ tends to increase with the number of links. One interesting result produced by the RDIRM is that representative features such as "swims," "water," "paws," "nestspot" and "meet" were found to have high relevance in their clusters. From these results, we can say that the relevance values estimated by the RDIRM indicate the popularity or representativeness of the objects. Consequently, we confirm that the RDIRM successfully discovered clear and major cluster structures by excluding structured noise in relational data.

3.5 Chapter Summary

In this chapter, we proposed a relevance dependent mechanism that enables biclustering models to distinguish between bicluster structure and structured noise. Then, we proposed a new probabilistic biclustering model called the Relevance Dependent Infinite Relational Model (RDIRM), that is suitable for analyzing relational data with structured noise. The RDIRM incorporates the relevance dependent mechanism, thereby

[3] For Enron, insufficient meta information for each employee is available. Therefore, for qualitative comparison, we only show the results for MovieLens and animal-feature.

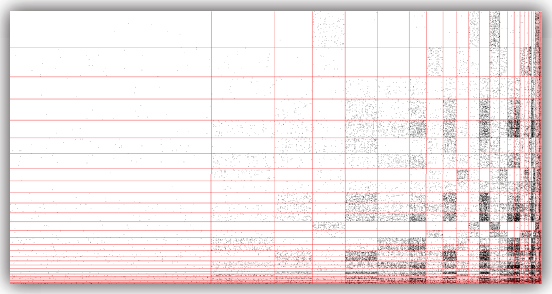
enabling the RDIRM to jointly estimate both a clear bicluster structure and structured noise.

Our experiments on synthetic datasets confirmed that the RDIRM can find proper clusters in relational data with structured noise, especially, in sparse or dense data. Moreover, our experiments on real-world datasets confirmed that the clusters obtained by the RDIRM represent major categories and that the estimated relevance parameters can be interpreted as measures of the popularity or representativeness of the objects.

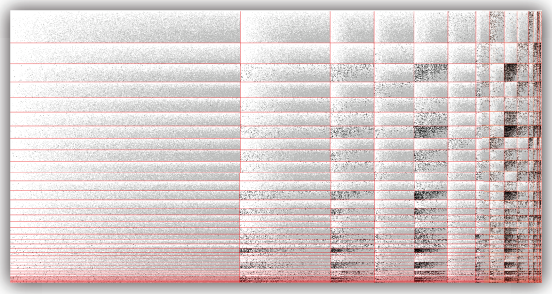
For future work, there are several promising extensions of the RDIRM can be considered. First, in our relevance dependent mechanism, structured noise underlying relationships is explained by only a single background probability η_0. Consequently, our mechanism can not explain the situation in which both non-link noise and link noise (i.e., spamming link) are present. Thus, a relevance dependent mechanism with multiple backgrounds can be expected to be helpful.

Second, the RDIRM includes the IRM for the clustering model; thus, objects are partitioned into non-overlapping clusters. However, mixed or multiple membership assumptions are appropriate for many real-world situations. Therefore, in the future, we plan to challenge the structured noise problem on more advanced relational models such as mixed (or multiple) membership models [1, 43], hierarchical structure models [57], and time-varying models [30, 19].

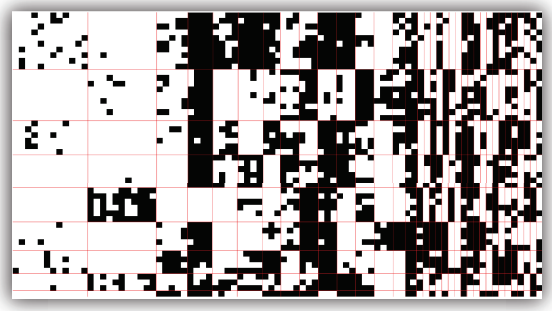
In addition, we are interested in developing a more efficient algorithm for the RDIRM that can handle large scale datasets. Although our MCMC based algorithm works efficiently, it is not sufficient for handling large scale data. Therefore, to make our RDIRM and its extensions available for business purposes, we plan to develop such a scalable algorithm in our future studies.



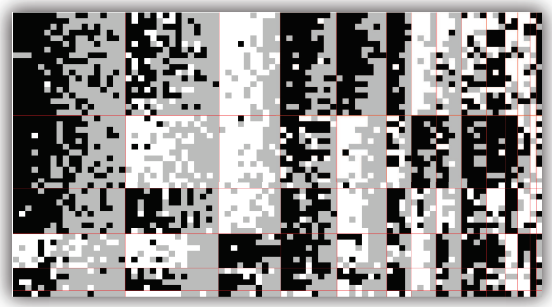
(a) MovieLens (IRM), TDLL = -0.135



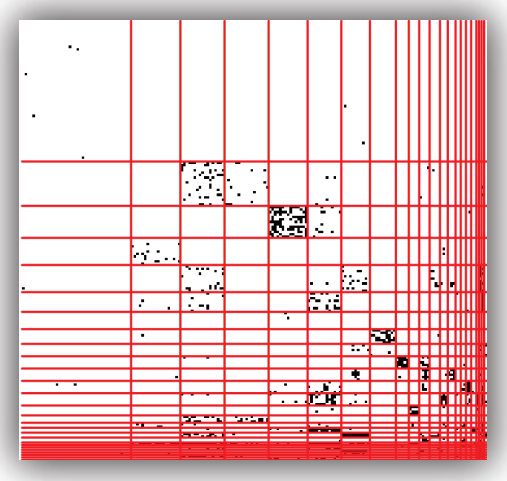
(b) MovieLens (RDIRM), TDLL = **-0.097**



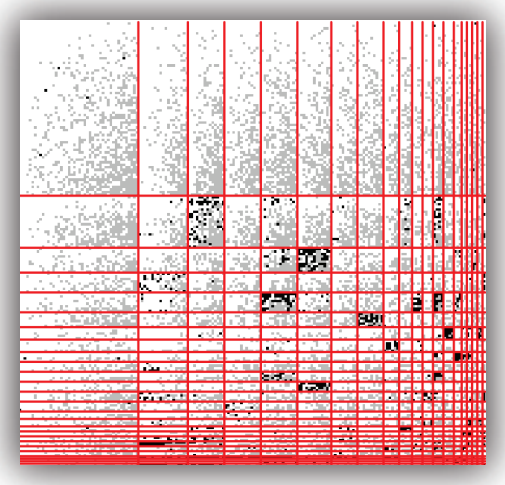
(c) animal-feature (IRM), TDLL = -0.393



(d) animal-feature (RDIRM), TDLL = **-0.213**

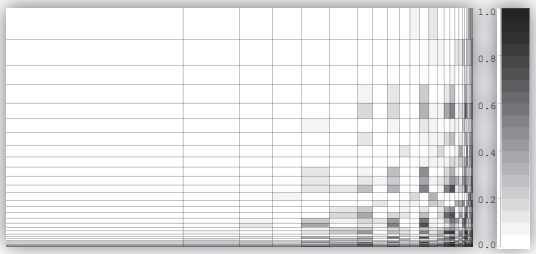


(e) Enron (IRM), TDLL = -0.055

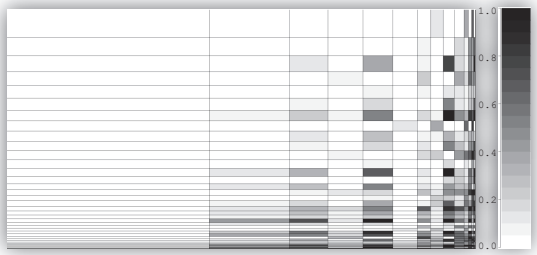


(f) Enron (RDIRM), TDLL = **-0.046**

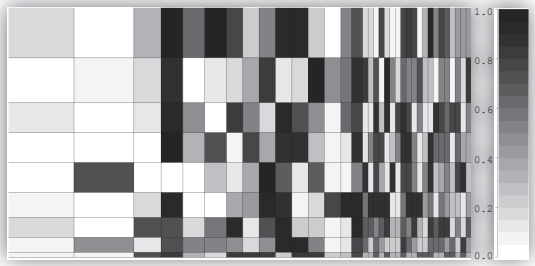
Figure 3.4: Clustering results for real-world datasets. Note that the objects within each cluster are sorted in descending order of estimated relevance values ${\rho_{1,i}}^{\mathrm{EAP}}$ and ${\rho_{2,j}}^{\mathrm{EAP}}$. "TDLL" is the computed test data log likelihood for each dataset.



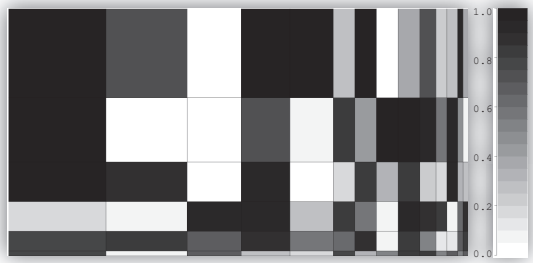
(a) MovieLens (IRM)



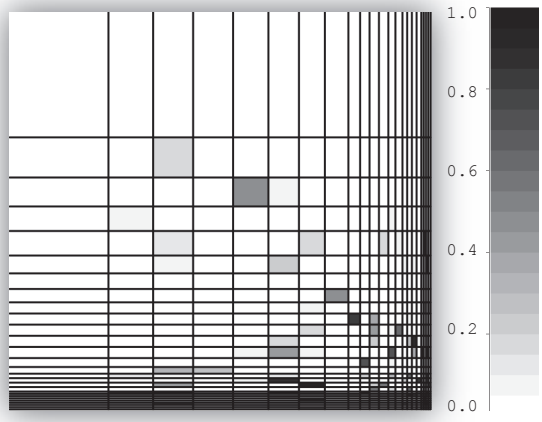
(b) MovieLens (RDIRM)



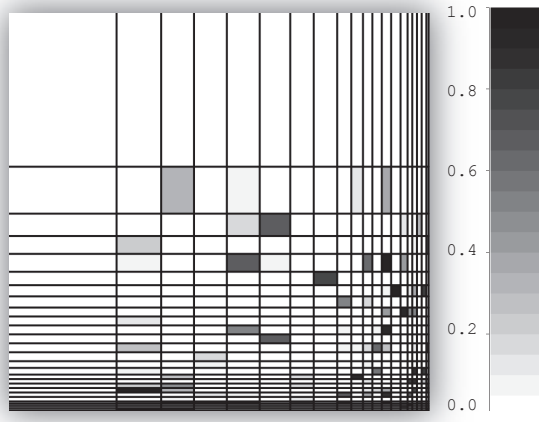
(c) animal-feature (IRM)



(d) animal-feature (RDIRM)



(e) Enron (IRM)



(f) Enron (RDIRM)

Figure 3.5: The estimated foreground link probabilities $\eta_{k,l}^{\mathrm{EAP}}$.

Table 3.3: Examples of clusters obtained by the RDIRM. The first column lists the object (Title/Feature). The second column lists the number of links related to the object (LNKS). The third column lists the estimated relevance (${\rho_{2,j}}^{\mathrm{EAP}}$). For the fourth column, we list the cluster indices obtained by the original IRM (ZIRM) to show that number of links affects the cluster assignments. The left and right side tables represent the MovieLens and animal-feature datasets, respectively.

Movie cluster 6			
Title	LNKS	${\rho_{2,j}}^{\mathrm{EAP}}$	ZIRM
Star Wars	501	0.9111	28
Return of the Jedi	379	0.5534	9
Independence Day	228	0.0921	25
Star Trek	220	0.1905	25

Movie cluster 7			
Title	LNKS	${\rho_{2,j}}^{\mathrm{EAP}}$	ZIRM
Silence of the Lambs	344	0.9132	26
Pulp Fiction	294	0.7598	26
Usual Suspects	232	0.6233	20
Alien	223	0.5164	20
Terminator	217	0.5608	20
Seven(Se7en)	167	0.3376	15

Movie Cluster 2			
Title	LNKS	${\rho_{2,j}}^{\mathrm{EAP}}$	ZIRM
W. W. & the C. F.	189	0.7196	27
Birdcage	154	0.4762	17
Truth About C. & D.	148	0.3386	17
Happy Gilmore	74	0.0360	2
Kingpin	73	0.1196	2

Feature cluster 1			
Feature	LNKS	${\rho_{2,j}}^{\mathrm{EAP}}$	ZIRM
swims	10	0.9808	2
water	10	0.9808	2
coastal	8	0.9231	2
arctic	9	0.8846	2
flippers	7	0.8077	2

Feature cluster 5			
Feature	LNKS	${\rho_{2,j}}^{\mathrm{EAP}}$	ZIRM
paws	19	0.9615	27
nestspot	31	0.9423	20
claws	19	0.9038	22
small	23	0.7885	21

Feature cluster 6			
Feature	LNKS	${\rho_{2,j}}^{\mathrm{EAP}}$	ZIRM
meat	20	0.9808	17
fierce	21	0.9231	17
hunter	17	0.8846	17
stalker	10	0.4808	16
scavenger	6	0.1538	1
flys	1	0.0769	1

Chapter 4

Relevance Modeling with Mixture Modeling

In this chapter, we generalize the relevance modeling in the RDIRM. By considering continuous relaxation of the Boolean function in the RDIRM, we propose a mixed-membership mechanism that contains all the Boolean functions as special cases. In the mixed-membership approach, we can resolve two critical limitations of relevance modeling in the RDIRM. First, in the mixed-membership mechanism, the form of the relaxed Boolean function can be automatically estimated from given data. Furthermore, in the mixed-membership mechanism, we can straightforwardly consider relevance values with three or more dimensions. Therefore, we can introduce multiple background distribution for considering different types of irrelevant objects. By incorporating the mixed-membership mechanism, we propose an extension of the RDIRM termed Multi-Layered IRM (MLIRM), which has two background distributions. The relevance parameters in the MLIRM can explain, not only passive objects with few links, but also spamming objects with extremely many links. Similar to the RDIRM, the posterior inference for the MLIRM can also be performed using collapsed Gibbs sampler. Experiments conducted using real-world datasets have confirmed that the

proposed model successfully extracts clear and interpretable cluster structures from real-world relational data.

4.1 Motivations

We propose a novel generative framework that captures a clear de-blurred cluster structure and object biases independently from blurred relational data. In the proposed framework, an observed link is drawn from a *mixture distribution of multiple layers*. The first layer is an abstract class of clustering models, such as the IRM. The other layers are uniform probabilities independent of the objects' cluster assignments. Then, the mixing ratio of each layer for a given pair of objects is controlled by the interaction of *bias parameters*, which are latent variables defined for each object. We propose a mechanism that describes the general form of interactions between biases and provide a hierarchical generative process for the mechanism. By estimating each layer, the bias parameters, and the form of interactions from given relational data, we can diminish the adverse effects of object biases and obtain a clear cluster structure. In addition, the estimated bias parameters enable us to extract representative objects strongly related to the underlying co-cluster structure. This is a great advantage of our framework, since one can easily understand the meaning of obtained clusters by inspecting only a few objects highly related to underlying co-cluster structure.

Since our multi-layered framework strictly follows the principles of hierarchical Bayesian modeling, we can incorporate the arbitrary Bayesian co-clustering models into the framework. By incorporating the IRM to the mixed-membership mechanism, we propose a new co-clustering model called the Multi-Layered Infinite Relational Model (MLIRM), which is a concrete instance of the proposed framework that incorporates the IRM. The MLIRM simultaneously estimates the object bias parameters and co-cluster structure underlying bias-corrected observations. Thanks to the property of the

embedded IRM, the MLIRM automatically estimates the number of clusters from the given data.

4.2 Multi-Layered Infinite Relational Model (MLIRM)

First, we propose the mixed-membership mechanism for modeling blurred relational data. Then, we propose a new generative model; i.e., the *Multi-Layered Infinite Relational Model* (MLIRM). Furthermore, we explain the relationships between the MLIRM and several recently proposed models including the IRM and RDIRM.

4.2.1 Mixed-Membership Mechanism

To capture the cluster structure and object biases independently, we propose a *mixed-membership mechanism*. In the proposed framework, we assume that each observation within relational data comes from a mixture distribution of three layers: a clustering layer $\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$, a background link layer $\eta_1 \in [0,1]$ and a background non-link layer $\eta_0 \in [0,1]$. The clustering layer is an arbitrary clustering model (e.g., IRM), which describes observations coming from the underlying cluster structure. The two background layers are probabilities independent from cluster assignment, which describe observations coming from object biases. The biased object means an object (e.g., user, item) with weak attribution to cluster structure. We consider two typical types of biased objects: a *spamming* object, which has extremely many links, and a *passive* object, which has few links (many non-links). To capture both links and non-links from biases, we introduce two background layers η_1 and η_0.

Here, we define a mechanism in which a mixture ratio of three layers for a pair of objects is controlled by the interaction of biases inherent to related objects. We introduce bias parameters $\boldsymbol{\theta}_{1,i} = (\theta_{1,i:\mathrm{fg}}, \theta_{1,i:\mathrm{bg1}}, \theta_{1,i:\mathrm{bg0}})^{\top}$ and $\boldsymbol{\theta}_{2,j} = (\theta_{2,j:\mathrm{fg}}, \theta_{2,j:\mathrm{bg1}}, \theta_{2,j:\mathrm{bg0}})^{\top}$ for each object, where fg, bg1, and bg0 indicate the clustering layer, the background link layer,

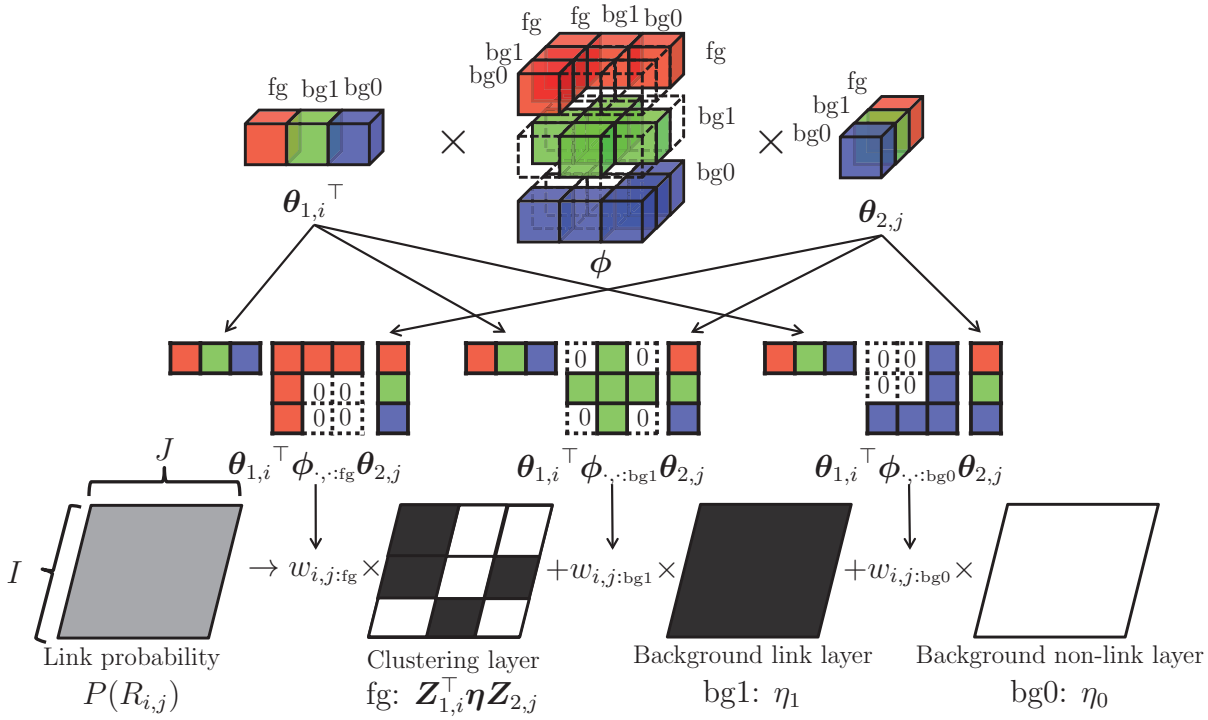


Figure 4.1: (Best viewed in color.) Diagram of the mixed-membership mechanism

and the background non-link layer, respectively. Each bias parameter is a probability vector whose dimension corresponds to the number of layers. These parameters control an object's relevance to each layer. Furthermore, let $\boldsymbol{w}_{i,j} = (w_{i,j:\mathrm{fg}}, w_{i,j:\mathrm{bg1}}, w_{i,j:\mathrm{bg0}})$ be the mixture ratio of each layer for an entry $R_{i,j}$. Then, we consider the interaction mechanism that transforms two bias parameters $\boldsymbol{\theta}_{1,i}$ and $\boldsymbol{\theta}_{2,j}$ into a mixture ratio $\boldsymbol{w}_{i,j}$. When considering the roles of object biases and the corresponding mixture ratio, it is important that there is a positive correlation between the two bias parameters $\boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j}$ and the corresponding mixture ratio $\boldsymbol{w}_{i,j}$. For example, when an bias parameter for the i-th row object leans to the foreground layer ($\theta_{1,i:\mathrm{fg}} \rightarrow 1.0$), the mixture ratio related to the i-th row object should also lean to the foreground ($w_{i,j:\mathrm{fg}} \rightarrow 1.0$). However, the form of transformation $f : \boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j} \rightarrow \boldsymbol{w}_{i,j}$ is non-trivial. Thus, we introduce a

constrained $3 \times 3 \times 3$ interaction weight $\boldsymbol{\phi}$ in order to describe the general form of the interaction between two biases. Using indices $s, t, u \in \{\text{fg}, \text{bg1}, \text{bg0}\}$, the mechanism is given as follows:

$$\begin{aligned} w_{i,j:u} =& {\boldsymbol{\theta}_{1,i}}^{\top} \boldsymbol{\phi}_{\cdot,\cdot:u} \boldsymbol{\theta}_{2,j} = \sum_s \sum_t \theta_{1,i:s} \theta_{2,j:t} \phi_{s,t:u} \\ & \text{s.t. } \textstyle\sum_{u'} \phi_{s,t:u'} = 1, \phi_{s,t:u} \in [0,1] \text{ and} \\ & \quad \phi_{s,t:u} = 0.0 \text{ if } (u \neq s \text{ and } u \neq t)\ \forall s, t, u, \end{aligned} \tag{4.1}$$

where $\boldsymbol{\phi}_{\cdot,\cdot:u}$ is a matrix that is a slice of $\boldsymbol{\phi}$ related to u. The constraints in Eq. (4.1) ensure a positive correlation between $\boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j}$ and $\boldsymbol{w}_{i,j}$. More intuitively, the constraints can be understood as follows:

- If row object i and column object j select the same layer according to $\boldsymbol{\theta}_{1,i}$ and $\boldsymbol{\theta}_{2,j}$, respectively, $R_{i,j}$ is generated from the layer with probability 1.0.
- If row object i and column object j select different layers from each other, one of the two layers is selected stochastically with a probability defined by the corresponding slice of $\boldsymbol{\phi}$. Finally, $R_{i,j}$ is generated from the selected layer.

Thus, in the proposed framework, link probability between two objects is given as follows:

$$\begin{aligned} & P(R_{i,j} = 1 \mid \boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j}, \eta_1, \eta_0, \boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j}, \boldsymbol{\phi}) \\ &= \begin{pmatrix} w_{i,j:\text{fg}} & w_{i,j:\text{bg1}} & w_{i,j:\text{bg0}} \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j} \\ \eta_1 \\ \eta_0 \end{pmatrix} \\ &= \left({\boldsymbol{\theta}_{1,i}}^{\top} \boldsymbol{\phi} \boldsymbol{\theta}_{2,j} \right)^{\top} \begin{pmatrix} \boldsymbol{Z}_{1,i}^{\top} \boldsymbol{\eta} \boldsymbol{Z}_{2,j} \\ \eta_1 \\ \eta_0 \end{pmatrix}, \end{aligned} \tag{4.2}$$

where ${\boldsymbol{\theta}_{1,i}}^{\top}\boldsymbol{\phi}\boldsymbol{\theta}_{2,j} = \left\{{\boldsymbol{\theta}_{1,i}}^{\top}\boldsymbol{\phi}_{\cdot,\cdot,u}\boldsymbol{\theta}_{2,j}\right\}_{u\in\{\mathrm{fg,bg1,bg0}\}}$.

To give the hierarchical generative model for the proposed framework, we use the extended definition of the Dirichlet distribution presented by Ferguson [17]. By taking the limit as any one of the Dirichlet parameters approaching zero, the corresponding random variables also degenerate to zero. Thus, a prior for $\boldsymbol{\phi}$ with the positive correlation constraint given by Eq. (4.1) can be constructed using Dirichlet distributions with constrained parameters. In summary, the hierarchical generative model of the multi-layered framework is as follows:

$$\boldsymbol{\theta}_{1,i} \,|\, \boldsymbol{\alpha}_1 \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_1),\ \boldsymbol{\theta}_{2,j} \,|\, \boldsymbol{\alpha}_2 \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_2), \tag{4.3}$$

$$\boldsymbol{\phi}_{s,t} \,|\, \boldsymbol{a}_{s,t} \sim \mathrm{Dirichlet}(\boldsymbol{a}_{s,t}) \quad \text{s.t. } a_{s,t:u} = 0.0 \text{ if } u \neq s \text{ and } u \neq t, \tag{4.4}$$

$$R_{i,j} \,|\, \boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}, \eta_1, \eta_0, \boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j}, \boldsymbol{\phi} \sim \mathrm{Bernoulli}\left(\left({\boldsymbol{\theta}_{1,i}}^{\top}\boldsymbol{\phi}\boldsymbol{\theta}_{2,j}\right)^{\top}\begin{pmatrix}\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}\\ \eta_1\\ \eta_0\end{pmatrix}\right), \tag{4.5}$$

where Dirichlet$(\cdot)$ is the Dirichlet distribution, and Bernoulli$(\cdot)$ is the Bernoulli distribution. Figure 4.1 presents a diagram of the proposed multi-layered framework. Now, we briefly review the above mentioned process. First, the bias parameters $\boldsymbol{\theta}_{1,i}$ and $\boldsymbol{\theta}_{2,j}$ for each object $O_{1,i}$ and $O_{2,j}$ are given from Dirichlet priors with parameters $\boldsymbol{\alpha}_1 = \{\alpha_{1,s}\}_{s\in\mathrm{fg,bg1,bg0}}$ and $\boldsymbol{\alpha}_2 = \{\alpha_{2,t}\}_{t\in\mathrm{fg,bg1,bg0}}$, respectively (Eq. (4.3)). Second, an interaction weight $\boldsymbol{\phi} = \{\boldsymbol{\phi}_{s,t}\}_{s,t\in\{\mathrm{fg,bg1,bg0}\}}$ is given by Dirichlet distributions with constrained hyperparameters $\boldsymbol{a}_{s,t}$ (Eq. (4.4)). Finally, the link $R_{i,j}$ is generated by a Bernoulli distribution (Eq. (4.5)).

4.2.2 The Multi-Layered Infinite Relational Model

Here, we propose a new generative model called the MLIRM, which is a concrete instance of the mixed-membership mechanism proposed in Section 4.2.1. In the MLIRM, the IRM is embedded as a prior for the clustering model $\boldsymbol{Z}_{1,i}^{\top}\boldsymbol{\eta}\boldsymbol{Z}_{2,j}$ given in Eq. (4.5). Therefore, in the IRM, the link probability of $R_{i,j}$ is given by $\eta_{k,l}$, where $z_{1,i} = k$ and $z_{2,j} = l$. To estimate the number of clusters automatically from the given data, the IRM uses the CRP as the prior distribution for $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. To summarize, the full description of the generative model for the MLIRM is as follows:

$$z_{1,i} \,|\, \gamma_1 \sim \mathrm{CRP}(\gamma_1),\ z_{2,j} \,|\, \gamma_2 \sim \mathrm{CRP}(\gamma_2), \tag{4.6}$$

$$\eta_{k,l} \,|\, \beta \sim \mathrm{Beta}(\beta, \beta), \tag{4.7}$$

$$\eta_1 \,|\, \beta \sim \mathrm{Beta}(\beta, \beta),\ \eta_0 = 1 - \eta_1, \tag{4.8}$$

$$\boldsymbol{\theta}_{1,i} \,|\, \boldsymbol{\alpha}_1 \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_1),\ \boldsymbol{\theta}_{2,j} \,|\, \boldsymbol{\alpha}_2 \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_2), \tag{4.9}$$

$$\boldsymbol{\phi}_{s,t} \,|\, \boldsymbol{a}_{s,t} \sim \mathrm{Dirichlet}(\boldsymbol{a}_{s,t}) \ \text{ s.t. } a_{s,t:u} = 0.0 \text{ if } u \neq s \text{ and } u \neq t, \tag{4.10}$$

$$R_{i,j} \,|\, z_{1,i}, z_{2,j}, \boldsymbol{\eta}, \eta_1, \eta_0, \boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j}, \boldsymbol{\phi} \sim \mathrm{Bernoulli}\left(\left({\boldsymbol{\theta}_{1,i}}^{\top} \boldsymbol{\phi} \boldsymbol{\theta}_{2,j} \right)^{\top} \begin{pmatrix} \eta_{z_{1,i}, z_{2,j}} \\ \eta_1 \\ \eta_0 \end{pmatrix} \right), \tag{4.11}$$

where $\mathrm{Beta}(\cdot,\cdot)$ is the beta distribution. Figure 4.2a shows a graphical representation for the MLIRM. Equations (4.6) and (4.7) are the embedded IRM for the clustering layer of the MLIRM. Note that γ_1 and γ_2 are the concentration parameters for the CRPs. Equation (4.8) defines the background layers. In the MLIRM, we define $\eta_1 = 1 - \eta_0$ in order to ensure that η_1 and η_0 capture irrelevant links and irrelevant non-links, respectively.

4.2.3 Relationships to Existing Relational Models

Here, we examine several recently proposed models that are closely related to the MLIRM, and show that the proposed MLIRM can be viewed as a generalization of

these models.

The IRM proposed by Kemp *et al.* [34] is one of the best-known clustering models that can account for a potentially infinite number of clusters in relational data. In the IRM, the link probability between $O_{1,i}$ and $O_{2,j}$ depends only on their non-overlapping cluster assignments $z_{1,i}$ and $z_{2,j}$. Then, the link $R_{i,j}$ is drawn as follows:

$$R_{i,j} \mid z_{1,i}, z_{2,j}, \boldsymbol{\eta} \sim \text{Bernoulli}\left(\eta_{z_{1,i},z_{2,j}}\right). \tag{4.12}$$

Note that the IRM is a special case of the proposed MLIRM when the bias parameters $\boldsymbol{\theta}_{1,i} = (\theta_{1,i:\text{fg}}, \theta_{1,i:\text{bg1}}, \theta_{1,i:\text{bg0}})^\top$ and $\boldsymbol{\theta}_{2,j} = (\theta_{2,j:\text{fg}}, \theta_{2,j:\text{bg1}}, \theta_{2,j:\text{bg0}})^\top$ are constrained to $(1, 0, 0)$; i.e., the IRM assumes relational data is generated only by the clustering layer of the MLIRM.

Clustering models considering the influences of the background layer were first discussed by Hoff [25, 26] relative to clustering biological sequences. For biclustering relational data, there have been several extensions of the IRM so that the background layer affect link probabilities. The Subset IRM (SIRM) proposed by Ishiguro *et al.* [33] and the Relevance Dependent IRM (RDIRM) proposed by Ohama *et al.* [48, 49] both consider a generative model in which link probability is a mixture distribution of a clustering layer $\eta_{k,l}$ and a background layer η_0. In the SIRM, binary variables $s_{1,i}, s_{2,j} \in \{0, 1\}$ are introduced to indicate whether each object is relevant to the underlying cluster structure. Then, a subset of $\boldsymbol{R}$, where $s_{1,i} \times s_{2,j} = 1$, is explained by the clustering model $\eta_{k,l}$, while the rest are explained by the background probability η_0. Specifically, in the SIRM, the link $R_{i,j}$ is drawn as follows:

$$\begin{aligned} &R_{i,j} \mid z_{1,i}, z_{2,j}, s_{1,i}, s_{2,j}, \boldsymbol{\eta}, \eta_0 \\ &\quad \sim \text{Bernoulli}\left(\begin{pmatrix} \mathbb{I}(s_{1,i} \times s_{2,j} = 1) \\ \mathbb{I}(s_{1,i} \times s_{2,j} = 0) \end{pmatrix}^\top \begin{pmatrix} \eta_{z_{1,i},z_{2,j}} \\ \eta_0 \end{pmatrix}\right), \end{aligned} \tag{4.13}$$

where $\mathbb{I}(\cdot)$ is 1 if the predicate holds and is zero otherwise. The RDIRM proposed by Ohama *et al.* relaxes the constraints of the SIRM. In the RDIRM, rather than $s_{1,i}$

and $s_{2,j}$ in the SIRM, continuous relevance parameters $\rho_{1,i}, \rho_{2,j} \in [0,1]$ are introduced; therefore, the RDIRM can estimate the confidence of an object being relevant to the underlying cluster structure. Furthermore, they introduced the static function $f : [0,1]^2 \to [0,1]$ to control the interaction between relevance parameters. Then, two versions of the RDIRM were proposed, the product model $f(\rho_{1,i}, \rho_{2,j}) = \rho_{1,i} \times \rho_{2,j}$ (RDIRM-prod) and the summation model $f(\rho_{1,i}, \rho_{2,j}) = 1-(1-\rho_{1,i})(1-\rho_{2,j})$ (RDIRM-sum). In general, the RDIRM generates a link as follows:

$$
\begin{aligned}
&R_{i,j} \mid z_{1,i}, z_{2,j}, \rho_{1,i}, \rho_{2,j}, \boldsymbol{\eta}, \eta_0 \\
&\quad \sim \text{Bernoulli}\left(\begin{pmatrix} f(\rho_{1,i}, \rho_{2,j}) \\ 1 - f(\rho_{1,i}, \rho_{2,j}) \end{pmatrix}^{\top} \begin{pmatrix} \eta_{z_{1,i}, z_{2,j}} \\ \eta_0 \end{pmatrix} \right). \qquad (4.14)
\end{aligned}
$$

The SIRM and RDIRM can be viewed as special cases of the MLIRM. The MLIRM is equivalent to the RDIRM in Eq. (4.14), if the bias parameters $\boldsymbol{\theta}_{1,i}$ and $\boldsymbol{\theta}_{2,j}$ are constrained to $\theta_{1,i:\text{bg1}} = \theta_{2,j:\text{bg1}} = 0$ and if the interaction tensor $\boldsymbol{\phi}$ is given statically. In addition to these constraints, the MLIRM is equivalent to the SIRM in Eq. (4.13) if $\boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j} \in \{(1,0,0),(0,0,1)\}$ and $\boldsymbol{\phi}_{\text{fg,bg0}} = \boldsymbol{\phi}_{\text{bg0,fg}} = (0,0,1)$.

Similar to the MLIRM, the RDIRM also considers that an entry $R_{i,j}$ is drawn from a mixture distribution of foreground and background layers. However, the RDIRM is too restrictive to capture bias values underlying real-world relationships. The RDIRM considers only the single background layer η_0. Therefore, the estimated value of η_0 tends to lean either on 0.0 or 1.0. Consequently, the RDIRM can only capture either passive or spamming objects. Furthermore, the RDIRM requires specifying the static function $f(\cdot,\cdot)$ appropriately for given data. In general, however, this is difficult as we often have no prior knowledge about given data. In contrast, our MLIRM can simultaneously capture both passive and spamming objects. In addition, the interaction function $\boldsymbol{\phi}$ can be estimated automatically from the data. To summarize, the advantages of the MLIRM are as follows.

- The MLIRM introduces two background layers η_1 and η_0 in order to accommodate relational data with both spamming and passive objects.
- The generative process for transformation $f : \boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,j} \rightarrow \boldsymbol{w}_{i,j}$ via $\boldsymbol{\phi}$ enables the interaction structure of two bias parameters to be estimated automatically.

4.3 Inference

In this section, we derive an efficient Gibbs sampler to perform posterior inference for the MLIRM.

4.3.1 Marginal Likelihood

Thanks to the conjugacy between MLIRM parameters and its prior distributions, $\boldsymbol{\eta}$, η_1, η_0, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\phi}$ can be fully marginalized out.

Introducing three auxiliary variables $r_{1,i\rightarrow j}, r_{2,j\rightarrow i}, r_{i,j} \in \{\text{fg}, \text{bg1}, \text{bg0}\}$, Eqs. (4.9)–(4.11) can be equivalently representaed by an augmented representation as follows:

$$z_{1,i} \,|\, \gamma_1 \sim \text{CRP}(\gamma_1),\ z_{2,j} \,|\, \gamma_2 \sim \text{CRP}(\gamma_2), \tag{4.15}$$

$$\eta_{k,l} \,|\, \beta \sim \text{Beta}(\beta, \beta), \tag{4.16}$$

$$\eta_1 \,|\, \beta \sim \text{Beta}(\beta, \beta),\ \eta_0 = 1 - \eta_1, \tag{4.17}$$

$$\boldsymbol{\theta}_{1,i} \,|\, \boldsymbol{\alpha}_1 \sim \text{Dirichlet}(\boldsymbol{\alpha}_1),\ \boldsymbol{\theta}_{2,j} \,|\, \boldsymbol{\alpha}_2 \sim \text{Dirichlet}(\boldsymbol{\alpha}_2), \tag{4.18}$$

$$\boldsymbol{\phi}_{s,t} \,|\, \boldsymbol{a}_{s,t} \sim \text{Dirichlet}(\boldsymbol{a}_{s,t}) \ \text{ s.t. } a_{s,t:u} = 0.0 \text{ if } u \neq s \text{ and } u \neq t, \tag{4.19}$$

$$r_{1,i\rightarrow j} \,|\, \boldsymbol{\theta}_{1,i} \sim \text{Categorical}(\boldsymbol{\theta}_{1,i}),\ r_{2,j\rightarrow i} \,|\, \boldsymbol{\theta}_{2,j} \sim \text{Categorical}(\boldsymbol{\theta}_{2,j}), \tag{4.20}$$

$$r_{i,j} \,|\, r_{1,i\rightarrow j}, r_{2,j\rightarrow i}, \boldsymbol{\phi} \sim \text{Categorical}(\boldsymbol{\phi}_{r_{1,i\rightarrow j}, r_{2,j\rightarrow i}}), \tag{4.21}$$

$$R_{i,j} \,|\, z_{1,i}, z_{2,j}, \boldsymbol{\eta}, \eta_1, \eta_0, r_{i,j} \sim \text{Bernoulli}\left(\begin{pmatrix} \mathbb{I}(r_{i,j} = \text{fg}) \\ \mathbb{I}(r_{i,j} = \text{bg1}) \\ \mathbb{I}(r_{i,j} = \text{bg0}) \end{pmatrix}^{\top} \begin{pmatrix} \eta_{z_{1,i}, z_{2,j}} \\ \eta_1 \\ \eta_0 \end{pmatrix} \right), \tag{4.22}$$

where Categorical$(\cdot)$ is the categorical distribution. The graphical representation of the augmented representation for the MLIRM is depicted in Fig. 4.2b.

Thanks to the conjugacy between categorical and Dirichlet distributions, marginalizing $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\phi}$ out, Eqs. (4.18)–(4.22) can be equivalently rewritten as follows:

$$r_{1,i\to j} \,|\, \boldsymbol{\alpha}_1 \sim \mathrm{DCM}(\boldsymbol{\alpha}_1),\ r_{2,j\to i} \,|\, \boldsymbol{\alpha}_2 \sim \mathrm{DCM}(\boldsymbol{\alpha}_2), \tag{4.23}$$

$$r_{i,j} \,|\, r_{1,i\to j} = s, r_{2,j\to i} = t, \boldsymbol{a}_{s,t} \sim \mathrm{DCM}(\boldsymbol{a}_{s,t}) \tag{4.24}$$

$$\text{s.t. } a_{s,t:u} = 0.0 \text{ if } u \neq s \text{ and } u \neq t,$$

$$R_{i,j} \,|\, z_{i,1}, z_{2,j}, \boldsymbol{\eta}, \eta_1, \eta_0, r_{i,j}$$

$$\sim \mathrm{Bernoulli}\left(\begin{pmatrix} \mathbb{I}(r_{i,j} = \mathrm{fg}) \\ \mathbb{I}(r_{i,j} = \mathrm{bg1}) \\ \mathbb{I}(r_{i,j} = \mathrm{bg0}) \end{pmatrix}^{\top} \begin{pmatrix} \eta_{z_{i,1},z_{2,j}} \\ \eta_1 \\ \eta_0 \end{pmatrix}\right), \tag{4.25}$$

where DCM$(\cdot)$ is the Dirichlet compound multinomial distribution. In Eqs. (4.23)–(4.25), the multinomial parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\phi}$ have been integrated out. Furthermore, given $r_{i,j}$, each observation $R_{i,j}$ is definitively assigned to one of the three layers. Thus, the Bernoulli parameters $\boldsymbol{\eta}$, η_1, and η_0 can also be integrated out. Consequently, we obtain a marginal representation of the MLIRM as depicted in Fig. 4.2c. Following the marginal representation, the closed-form marginal likelihood for the MLIRM is described as follows:

$$\begin{aligned} &P(\boldsymbol{R}, \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}, \boldsymbol{r}_1, \boldsymbol{r}_2) \\ &\quad = P(\boldsymbol{R} \,|\, \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}) \times P(\boldsymbol{r} \,|\, \boldsymbol{r}_1, \boldsymbol{r}_2) \\ &\qquad \times P(\boldsymbol{r}_1) \times P(\boldsymbol{r}_2) \times P(\boldsymbol{z}_1) \times P(\boldsymbol{z}_2), \end{aligned} \tag{4.26}$$

where the terms on the right hand side of Eq. (4.26) can be derived as described below.

First, thanks to the conjugacy between the beta and Bernoulli distributions, the

first term on the right hand side of Eq. (4.26) is derived as follows:

$$
\begin{aligned}
&P(\boldsymbol{R} \,|\, \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}) \\
&\quad = \iint P(\boldsymbol{R} \,|\, \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}, \boldsymbol{\eta}, \eta_1) P(\boldsymbol{\eta}) P(\eta_1) \mathrm{d}\boldsymbol{\eta} \mathrm{d}\eta_1 \\
&\quad = \frac{B(m_{\mathrm{bg1}} + \overline{m}_{\mathrm{bg0}} + \beta, m_{\mathrm{bg0}} + \overline{m}_{\mathrm{bg1}} + \beta)}{B(\beta, \beta)} \\
&\qquad \times \prod_k \prod_l \frac{B(m_{k,l:\mathrm{fg}} + \beta, \overline{m}_{k,l:\mathrm{fg}} + \beta)}{B(\beta, \beta)},
\end{aligned} \tag{4.27}
$$

where $B(\cdot, \cdot)$ denotes the beta function. The symbols m and $\overline{m}$ are the number of links and non-links, and are computed as follows:

$$
m_{\mathrm{bg1}} = \sum_{i \in T_1} \sum_{j \in T_2} R_{i,j} \mathbb{I}(r_{i,j} = \mathrm{bg1}), \tag{4.28}
$$

$$
m_{\mathrm{bg0}} = \sum_{i \in T_1} \sum_{j \in T_2} R_{i,j} \mathbb{I}(r_{i,j} = \mathrm{bg0}), \tag{4.29}
$$

$$
m_{k,l:\mathrm{fg}} = \sum_{i \in T_1} \sum_{j \in T_2} R_{i,j} \mathbb{I}(r_{i,j} = \mathrm{fg}) \mathbb{I}(z_{1,i} = k) \mathbb{I}(z_{2,j} = l), \tag{4.30}
$$

$$
\overline{m}_{\mathrm{bg1}} = \sum_{i \in T_1} \sum_{j \in T_2} (1 - R_{i,j}) \mathbb{I}(r_{i,j} = \mathrm{bg1}), \tag{4.31}
$$

$$
\overline{m}_{\mathrm{bg0}} = \sum_{i \in T_1} \sum_{j \in T_2} (1 - R_{i,j}) \mathbb{I}(r_{i,j} = \mathrm{bg0}), \tag{4.32}
$$

$$
\overline{m}_{k,l:\mathrm{fg}} = \sum_{i \in T_1} \sum_{j \in T_2} (1 - R_{i,j}) \mathbb{I}(r_{i,j} = \mathrm{fg}) \mathbb{I}(z_{1,i} = k) \mathbb{I}(z_{2,j} = l). \tag{4.33}
$$

Second, the terms related to the DCM distribution in Eq. (4.26) are derived as

follows:

$$
\begin{aligned}
&P(\boldsymbol{r} \mid \boldsymbol{r}_1, \boldsymbol{r}_2) \\
&\quad = \prod_s \prod_t \frac{\Gamma(\sum_u a_{s,t:u})}{\Gamma(\sum_u (a_{s,t:u} + n_{s,t:u}))} \\
&\qquad \times \prod_s \prod_t \prod_{u \in \{s,t\}} \prod_u \frac{\Gamma(a_{s,t:u} + n_{s,t:u})}{\Gamma(a_{s,t:u})}, \qquad (4.34)
\end{aligned}
$$

$$
P(\boldsymbol{r}_1) = \left(\frac{\Gamma(\sum_s \alpha_{1,s})}{\Gamma(\sum_s \alpha_{1,s} + J)} \right)^I \prod_{i \in T_1} \prod_s \frac{\Gamma(\alpha_{1,s} + n_{1,i:s})}{\Gamma(\alpha_{1,s})}, \qquad (4.35)
$$

$$
P(\boldsymbol{r}_2) = \left(\frac{\Gamma(\sum_t \alpha_{2,t})}{\Gamma(\sum_t \alpha_{2,t} + I)} \right)^J \prod_{j \in T_2} \prod_t \frac{\Gamma(\alpha_{2,t} + n_{2,j:t})}{\Gamma(\alpha_{2,t})}, \qquad (4.36)
$$

where $\Gamma(\cdot)$ is the gamma function. The symbols n, n_1, and n_2 denote the counts defined for the auxiliary variables $\boldsymbol{r}$, $\boldsymbol{r}_1$, and $\boldsymbol{r}_2$, respectively, and are computed as follows:

$$
n_{s,t:u} = \sum_{i \in T_1} \sum_{j \in T_2} \mathbb{I}(r_{1,i \to j} = s) \mathbb{I}(r_{2,j \to i} = t) \mathbb{I}(r_{i,j} = u), \qquad (4.37)
$$

$$
n_{1,i:s} = \sum_{j \in T_2} \mathbb{I}(r_{1,i \to j} = s), \quad n_{2,j:t} = \sum_{i \in T_1} \mathbb{I}(r_{2,j \to i} = t). \qquad (4.38)
$$

Finally, the cluster assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ follow the CRP. Therefore, we obtain

$$
P(\boldsymbol{z}_1) = \gamma_1^K \frac{\Gamma(\gamma_1) \prod_k \Gamma(m_{1,k})}{\Gamma(I + \gamma_1)}, \qquad (4.39)
$$

$$
P(\boldsymbol{z}_2) = \gamma_2^L \frac{\Gamma(\gamma_2) \prod_l \Gamma(m_{2,l})}{\Gamma(J + \gamma_2)}, \qquad (4.40)
$$

where $m_{1,k} = \sum_i \mathbb{I}(z_{1,i} = k)$ and $m_{2,l} = \sum_j \mathbb{I}(z_{2,j} = l)$.

4.3.2 Posterior Inference

Since the closed-form marginal likelihood for the MLIRM is now available, the posterior inference for the MLIRM can be performed efficiently using collapsed inference methods [41, 70]. In this paper, we use the collapsed Gibbs sampler [41] to infer the

MLIRM because the algorithm guarantees asymptotic convergence to the true posterior by drawing infinitely many samples.

As the parameters for the MLIRM have been marginalized out, the only variables we have to estimate are $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}$.

Sampling Cluster Assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$

$z_{1,i}$ and $z_{2,j}$ can be sampled in the same way; thus, here, we concentrate only on $z_{1,i}$. Given $\boldsymbol{r}$, the cluster assignments depend only on the subset of observations, where $r_{i,j} = \text{fg}$. Therefore, the conditional posterior for $z_{1,i} = k^*$ is derived from Eq. (4.26) as follows:

$$P(z_{1,i} = k^* \mid \boldsymbol{z}_{1,-i}, \boldsymbol{z}_2, \boldsymbol{r}, \boldsymbol{R}) \propto \begin{cases} m_{1,-i,k^*} \times \prod_l \frac{B(m^{+i}_{k^*,l:\text{fg}}+\beta, \overline{m}^{+i}_{k^*,l:\text{fg}}+\beta)}{B(m^{-i}_{k^*,l:\text{fg}}+\beta, \overline{m}^{-i}_{k^*,l:\text{fg}}+\beta)} & m_{1,-i,k^*} > 0, \\ \gamma_1 \times \prod_l \frac{B(m^{+i}_{k^*,l:\text{fg}}+\beta, \overline{m}^{+i}_{k^*,l:\text{fg}}+\beta)}{B(\beta,\beta)} & m_{1,-i,k^*} = 0, \end{cases} \tag{4.41}$$

where $\boldsymbol{z}_{1,-i}$ denotes the cluster assignments for all objects in T_1 excluding $O_{1,i}$ and $m_{1,-i,k^*}$ is the number of objects assigned to cluster k^* excluding $O_{1,i}$. The symbols m_{fg} and $\overline{m}_{\text{fg}}$ with superscripts are computed as follows:

$$m^{-i}_{k^*,l:\text{fg}} = \sum_{x \neq i} \sum_{j \in T_2} R_{x,j} \mathbb{I}(r_{x,j}=\text{fg}) \mathbb{I}(z_{1,x}=k^*) \mathbb{I}(z_{2,j}=l), \tag{4.42}$$

$$\overline{m}^{-i}_{k^*,l:\text{fg}} = \sum_{x \neq i} \sum_{j \in T_2} (1-R_{x,j}) \mathbb{I}(r_{x,j}=\text{fg}) \mathbb{I}(z_{1,x}=k^*) \mathbb{I}(z_{2,j}=l), \tag{4.43}$$

$$m^{+i}_{k^*,l:\text{fg}} = m^{-i}_{k^*,l:\text{fg}} + \sum_{j \in T_2} R_{i,j} \mathbb{I}(z_{2,j} = l), \tag{4.44}$$

$$\overline{m}^{+i}_{k^*,l:\text{fg}} = \overline{m}^{-i}_{k^*,l:\text{fg}} + \sum_{j \in T_2} (1 - R_{i,j}) \mathbb{I}(z_{2,j} = l). \tag{4.45}$$

Sampling Auxiliary Variables $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}$

The naive sampler of $r_{1,i\to j}$, $r_{2,j\to i}$, and $r_{i,j}$ can be derived in a straightforward manner. However, sampling these variables one after the other causes slow mixing of the Markov

chain because these variables are highly correlated. Therefore, for efficient mixing, we group these variables and update them simultaneously. Here, let $\boldsymbol{r}^{\mathrm{g}}_{i,j}$ be the grouped variables $\{r_{1,i\to j}, r_{2,j\to i}, r_{i,j}\}$. Then, the conditional posterior is given as follows:

$$\begin{aligned}
&P(\boldsymbol{r}^{\mathrm{g}}_{i,j}{=}\{s^*, t^*, u^*\} \mid \boldsymbol{R}, \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}_{-(i,j)}, \boldsymbol{r}_{1,-(i\to j)}, \boldsymbol{r}_{2,-(j\to i)}) \\
&\quad \propto P(R_{i,j} \mid \boldsymbol{R}_{-(i,j)}, \boldsymbol{z}_1, \boldsymbol{z}_2, r_{i,j}{=}u^*, \boldsymbol{r}_{-(i,j)}) \\
&\qquad \times P(r_{i,j}{=}u^* \mid \boldsymbol{r}_{-(i,j)}, r_{1,i\to j}{=}s^*, \boldsymbol{r}_{1,-(i\to j)}, r_{2,j\to i}{=}t^*, \boldsymbol{r}_{2,-(j\to i)}) \\
&\qquad \times P(r_{1,i\to j}{=}s^* \mid \boldsymbol{r}_{1,-(i\to j)}) \\
&\qquad \times P(r_{2,j\to i}{=}t^* \mid \boldsymbol{r}_{2,-(j\to i)}),
\end{aligned} \tag{4.46}$$

where $\boldsymbol{r}_{1,-(i\to j)}$ and $\boldsymbol{r}_{2,-(j\to i)}$ denote the entire set of $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ excluding $r_{1,i\to j}$ and $r_{2,j\to i}$, respectively. Similarly, $\boldsymbol{r}_{-(i,j)}$ denotes the entire set of $\boldsymbol{r}$ without $r_{i,j}$. The terms on the right hand side of Eq. (4.46) are computed as follows:

$$\begin{aligned}
&P(R_{i,j} \mid \boldsymbol{R}_{-(i,j)}, \boldsymbol{z}_1, \boldsymbol{z}_2, r_{i,j} = u^*, \boldsymbol{r}_{-(i,j)}) \\
&= \begin{cases}
\frac{(m^{-(i,j)}_{k,l:\mathrm{fg}}+\beta)^{R_{i,j}}(\overline{m}^{-(i,j)}_{k,l:\mathrm{fg}}+\beta)^{\overline{R}_{i,j}}}{m^{-(i,j)}_{k,l:\mathrm{fg}}+\overline{m}^{-(i,j)}_{k,l:\mathrm{fg}}+2\beta} & u^*{=}\mathrm{fg}, \\
\frac{(m^{-(i,j)}_{\mathrm{bg1}}+\overline{m}^{-(i,j)}_{\mathrm{bg0}}+\beta)^{R_{i,j}}(\overline{m}^{-(i,j)}_{\mathrm{bg1}}+m^{-(i,j)}_{\mathrm{bg0}}+\beta)^{\overline{R}_{i,j}}}{m^{-(i,j)}_{\mathrm{bg1}}+\overline{m}^{-(i,j)}_{\mathrm{bg0}}+\overline{m}^{-(i,j)}_{\mathrm{bg1}}+m^{-(i,j)}_{\mathrm{bg0}}+2\beta} & u^*{=}\mathrm{bg1}, \\
\frac{(m^{-(i,j)}_{\mathrm{bg0}}+\overline{m}^{-(i,j)}_{\mathrm{bg1}}+\beta)^{R_{i,j}}(\overline{m}^{-(i,j)}_{\mathrm{bg0}}+m^{-(i,j)}_{\mathrm{bg1}}+\beta)^{\overline{R}_{i,j}}}{m^{-(i,j)}_{\mathrm{bg0}}+\overline{m}^{-(i,j)}_{\mathrm{bg1}}+\overline{m}^{-(i,j)}_{\mathrm{bg0}}+m^{-(i,j)}_{\mathrm{bg1}}+2\beta} & u^*{=}\mathrm{bg0},
\end{cases}
\end{aligned} \tag{4.47}$$

$$\begin{aligned}
&P(r_{i,j}{=}u^* \mid \boldsymbol{r}_{-(i,j)}, r_{1,i\to j}{=}s^*, \boldsymbol{r}_{1,-(i\to j)}, r_{2,j\to i}{=}t^*, \boldsymbol{r}_{2,-(j\to i)}) \\
&\quad = \frac{n^{-(i,j)}_{s^*,t^*:u^*} + a_{s^*,t^*:u^*}}{\sum_u (n^{-(i,j)}_{s^*,t^*:u} + a_{s^*,t^*:u})},
\end{aligned} \tag{4.48}$$

$$P(r_{1,i\to j} = s^* \mid \boldsymbol{r}_{1,-(i\to j)}) \propto \alpha_{1,s^*} + \sum_{y\in T_2: y\neq j} \mathbb{I}(r_{1,i\to y} = s^*), \tag{4.49}$$

$$P(r_{2,j\to i} = t^* \mid \boldsymbol{r}_{2,-(j\to i)}) \propto \alpha_{2,t^*} + \sum_{x\in T_1: x\neq i} \mathbb{I}(r_{2,j\to x} = t^*), \tag{4.50}$$

where the related counts are computed as follows:

$$m_{k,l:\mathrm{fg}}^{-(i,j)} = m_{k,l:\mathrm{fg}} - R_{i,j}\mathbb{I}(r_{i,j} = \mathrm{fg})\mathbb{I}(z_{1,i} = k)\mathbb{I}(z_{2,j} = l), \tag{4.51}$$

$$m_{\mathrm{bg1}}^{-(i,j)} = m_{\mathrm{bg1}} - R_{i,j}\mathbb{I}(r_{i,j} = \mathrm{bg1}), \tag{4.52}$$

$$m_{\mathrm{bg0}}^{-(i,j)} = m_{\mathrm{bg0}} - R_{i,j}\mathbb{I}(r_{i,j} = \mathrm{bg0}), \tag{4.53}$$

$$\overline{m}_{k,l:\mathrm{fg}}^{-(i,j)} = m_{k,l:\mathrm{fg}} - (1 - R_{i,j})\mathbb{I}(z_{1,i} = k)\mathbb{I}(z_{2,j} = l), \tag{4.54}$$

$$\overline{m}_{\mathrm{bg1}}^{-(i,j)} = \overline{m}_{\mathrm{bg1}} - (1 - R_{i,j})\mathbb{I}(r_{i,j} = \mathrm{bg1}), \tag{4.55}$$

$$\overline{m}_{\mathrm{bg0}}^{-(i,j)} = \overline{m}_{\mathrm{bg0}} - (1 - R_{i,j})\mathbb{I}(r_{i,j} = \mathrm{bg0}), \tag{4.56}$$

$$n_{s,t:u}^{-(i,j)} = n_{s,t:u} - \mathbb{I}(r_{1,i\to j} = s)\mathbb{I}(r_{2,j\to i} = t)\mathbb{I}(r_{i,j} = u). \tag{4.57}$$

4.3.3 Estimating Hyperparameters

In general, the hyperparameters of a statistical model should be tuned carefully in order to obtain a better solution. For the hyperparameters of the MLIRM, we derive posterior samplers using *data augmentation* [16, 74, 69, 45] techniques.

Data Augmentation

Let us consider two situations that have the following probability densities:

$$P(N \mid U) \propto \frac{\Gamma(U)}{\Gamma(U + N)}, \tag{4.58}$$

$$P(N' \mid U') \propto \frac{\Gamma(U' + N')}{\Gamma(U')}, \tag{4.59}$$

where N, N' are positive integer random variables and U, U' are positive real random variables, respectively. Unfortunately, in these case, we cannot derive straightforward posterior Gibbs samplers for U and U' because conjugate priors for Eqs. (4.58) and (4.59) have not been developed so far.

To over come the above-mentioned difficulties, data augmentation techniques consider expanded joint probabilities over target and auxiliary variables. Let us denote

by $\mathrm{Gam}(e_0, f_0)$ the gamma prior with shape parameter e_0 and rate parameter f_0; i.e., $P(\lambda \,|\, e_0, f_0) = \lambda^{e_0-1} e^{-f_0 \lambda} / G(e_0, f_0)$ where $G(e_0, f_0) = \Gamma(e_0)/{f_0}^{e_0}$. The key strategies are to use the following expansions:

$$\frac{\Gamma(U)}{\Gamma(U+N)} = \frac{1}{\Gamma(N)} \int_0^1 p^{U-1}(1-p)^{N-1} \mathrm{d}p, \tag{4.60}$$

$$\frac{\Gamma(U'+N')}{\Gamma(U')} = \sum_{q=1}^{N'} S(N', q) U'^q, \tag{4.61}$$

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

By expanding Eq. (4.58) using Eq. (4.60), the joint distribution over N and p given U is described as

$$P(N, p \,|\, U) \propto \frac{1}{\Gamma(N)} p^{U-1}(1-p)^{N-1}. \tag{4.62}$$

Therefore, a random sample from posterior $P(p \,|\, N, U)$ can be obtained as

$$p \,|\, N, U \sim \mathrm{Beta}(U, N), \tag{4.63}$$

because $P(p \,|\, N, U) = P(N, p \,|\, U)/P(N \,|\, U) = p^{U-1}(1-p)^{N-1}/B(U, N)$. Given p and assuming gamma prior as $U \sim \mathrm{Gamma}(e_0, f_0)$, the posterior for U is given by

$$P(U \,|\, N, p) \propto P(p \,|\, N, U) P(U) \propto e^{U \ln p} \times U^{e_0-1} e^{-f_0 U}. \tag{4.64}$$

Consequently, posterior sampling for U can be performed as

$$U \,|\, N, p \sim \mathrm{Gamma}(e_0, f_0 - \ln p). \tag{4.65}$$

Similarly, by expanding Eq. (4.59) using Eq. (4.61), the joint distribution over N and q, given U, is described as

$$P(N', q \,|\, U') \propto S(N', q) U'^q, \tag{4.66}$$

where the posterior for q follows an Antoniak distribution [3] as

$$q \,|\, N', U' \sim \mathrm{Antoniak}(N', U'). \tag{4.67}$$

The Antoniak distribution (also called the Chinese Restaurant Table distribution [75]) is the distribution of the number of occupied tables if N' customers are assigned to one of infinite tables with CRP(U'), and is sampled as $q \sim \sum_{w=1}^{N'} \text{Bernoulli}(U'/(U'+w-1))$. Given q and assuming gamma prior as $U' \sim \text{Gamma}(e_0, f_0)$, the posterior for U' is given by

$$P(U' \mid N', q) \propto P(q \mid N', U')P(U') \propto U'^{q} \times U'^{e_0-1} e^{-f_0 U'}. \tag{4.68}$$

Consequently, the posterior sampling for U' can be performed as

$$U' \mid N', q \sim \text{Gamma}(e_0 + q, f_0). \tag{4.69}$$

Sampling Hyperparameters

In this section, we show that posterior samplers for all hyperparameters (i.e., γ_1, γ_2, β, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{a}$) can be derived using the data augmentation techniques we introduced in Section 4.3.3.

Since a beta function is equivalently rewritten as $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, Eq. (4.27) can be equivalently rewritten as follows:

$$\begin{aligned}
&P(\boldsymbol{R} \mid \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}) \\
&= \frac{\Gamma(2\beta)}{\Gamma(m_{\text{bg1}} + \overline{m}_{\text{bg0}} + m_{\text{bg0}} + \overline{m}_{\text{bg1}} + 2\beta)} \\
&\quad \times \frac{\Gamma(m_{\text{bg1}} + \overline{m}_{\text{bg0}} + \beta)}{\Gamma(\beta)} \times \frac{\Gamma(m_{\text{bg0}} + \overline{m}_{\text{bg1}} + \beta)}{\Gamma(\beta)} \\
&\quad \times \prod_k \prod_l \left\{ \frac{\Gamma(2\beta)}{\Gamma(m_{k,l:\text{fg}} + \overline{m}_{k,l:\text{fg}} + 2\beta)} \times \frac{\Gamma(m_{k,l:\text{fg}} + \beta)}{\Gamma(\beta)} \times \frac{\Gamma(m_{k,l:\text{fg}} + \beta)}{\Gamma(\beta)} \right\}.
\end{aligned} \tag{4.70}$$

Therefore, by expanding Eq. (4.70) using Eqs. (4.60) and (4.61), we can obtain a joint distribution over β and several auxiliary variables, where posterior samplers for the

auxiliary variables are derived as follows:

$$p_{\mathrm{bg}} \sim \mathrm{Beta}(2\beta, m_{\mathrm{bg1}} + \overline{m}_{\mathrm{bg0}} + m_{\mathrm{bg0}} + \overline{m}_{\mathrm{bg1}}), \tag{4.71}$$

$$q_{\mathrm{bg}} \sim \mathrm{Antoniak}(m_{\mathrm{bg1}} + \overline{m}_{\mathrm{bg0}}, \beta), \tag{4.72}$$

$$\overline{q}_{\mathrm{bg}} \sim \mathrm{Antoniak}(m_{\mathrm{bg0}} + \overline{m}_{\mathrm{bg1}}, \beta), \tag{4.73}$$

$$p_{k,l} \sim \mathrm{Beta}(2\beta, m_{k,l:\mathrm{fg}} + \overline{m}_{k,l:\mathrm{fg}}), \tag{4.74}$$

$$q_{k,l} \sim \mathrm{Antoniak}(m_{k,l:\mathrm{fg}}, \beta), \tag{4.75}$$

$$\overline{q}_{k,l} \sim \mathrm{Antoniak}(\overline{m}_{k,l:\mathrm{fg}}, \beta). \tag{4.76}$$

Consequently, assuming the prior as $\beta \sim \mathrm{Gamma}(e_0, f_0)$, β can be updated as

$$\begin{aligned}\beta \,|\, - \sim \mathrm{Gamma}(&e_0 + q_{\mathrm{bg}} + \overline{q}_{\mathrm{bg}} + \textstyle\sum_k \sum_l (q_{k,l} + \overline{q}_{k,l}),\\ &f_0 - 2\ln p_{\mathrm{bg}} - \textstyle\sum_k \sum_l \ln p_{k,l}),\end{aligned} \tag{4.77}$$

where $\beta \,|\, -$ denotes a posterior sample of β given all the remaining variables.

For $\boldsymbol{a}$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, γ_1, and γ_2, posterior samplers can be straightforwardly derived by expanding Eqs. (4.34), (4.35), (4.36), (4.39), and (4.40) using Eqs. (4.60) and (4.61). Consequently, these hyperparameters can be updated as follows:

$$p_{s,t} \sim \mathrm{Beta}(\textstyle\sum_u a_{s,t:u}, \sum_u n_{s,t:u}), \tag{4.78}$$

$$q_{s,t:u} \sim \mathrm{Antoniak}(n_{s,t:u}, a_{s,t:u}), \tag{4.79}$$

$$a_{s,t:u} \,|\, - \sim \mathrm{Gamma}(e_0 + q_{s,t:u}, f_0 - \ln p_{s,t}), \tag{4.80}$$

$$p \sim \mathrm{Beta}(\textstyle\sum_s \alpha_{1,s}, J), \tag{4.81}$$

$$q_{i,s} \sim \mathrm{Antoniak}(n_{1,i:s}, \alpha_{1,s}), \tag{4.82}$$

$$\alpha_{1,s} \,|\, - \sim \mathrm{Gamma}(e_0 + \textstyle\sum_i q_{i,s}, f_0 - I \ln p), \tag{4.83}$$

$$p \sim \mathrm{Beta}(\gamma_1, I), \gamma_1 \,|\, - \sim \mathrm{Gamma}(e_0 + K, f_0 - \ln p). \tag{4.84}$$

Note that posterior samplers for $\boldsymbol{\alpha}_2$ and γ_2 are omitted because these can be sampled in the same way as $\boldsymbol{\alpha}_1$ and γ_1, respectively.

Algorithm 1 Collapsed Gibbs inference for the MLIRM

$\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r} \leftarrow$ Initialize latent variables
repeat
  for $i = 1$ to I **do**
    $z_{1,i} \leftarrow$ Random sample using Eq. (4.41)
  end for
  for $j = 1$ to J **do**
    $z_{2,j} \leftarrow$ Random sample in same manner as $z_{1,i}$
  end for
  for $i = 1$ to I **do**
    for $j = 1$ to J **do**
      $\{r_{1,i\to j}, r_{2,j\to i}, r_{i,j}\} \leftarrow$ Random sample using Eq. (4.46)
    end for
  end for
  Update hyperparameters according to Section 4.3.3
until convergence of marginal likelihood Eq. (4.26)

4.3.4 Pseudocode for Performing Posterior Inference

Using the posterior samplers derived in Sections 4.3.2 and 4.3.3, the posterior inference for the MLIRM can be completely performed by a closed-form Gibbs sampling algorithm. The pseudocode for the inference algorithm is summarized as Algorithm 1.

4.3.5 Computational Efficiency

Here, we briefly discuss the computational cost of the MLIRM and related models. As evident from the form of Eq. (4.41), posterior update for a cluster assignment $z_{1,i}$ requires $O(KL)$ computation. Therefore, updating $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ requires $O((I +$

$J)KL$) computation. Additionally, the MLIRM requires $O(IJ)$ computation to update auxiliary variables $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}$. Consequently, collapsed Gibbs sampling for the MLIRM roughly requires an $O((I + J)KL + IJ)$ computation, which is the same as that for the RDIRM. For the IRM and SIRM, collapsed Gibbs samplers require an $O((I+J)KL)$ computation [51, 53]. Therefore, the MLIRM requires more computation than the IRM or SIRM. However, the computational efficiency of the generative models depends very much on the choice of inference algorithm. Therefore, evaluating the computational efficiency of these models using more modern inference algorithms [70, 32, 11] is one of the promising directions for future work.

4.4 Experiments

In this section, we present the experimental results using a toy synthetic dataset and several real-world datasets, i.e., "Animal" [55], "Enron" [35], and "MovieLens[1]". In all experiments, we fit the hyperparameters of the MLIRM using the samplers we derived in Section 4.3.2. In addition, the hyperparameters for the conventional models are also estimated using Gibbs samplers, which can be derived via same data augmentation techniques. Note that we set $e_0 = f_0 = 1.0$ for all models in all experiments discussed in this paper.

4.4.1 Synthetic Data

First, we explored the ability of the proposed MLIRM to recover the underlying cluster structure using synthetic data. Figure 4.3a shows the hand-constructed synthetic data used in this experiment. As can be seen in Fig. 4.3a, there are many biased objects with extremely many links or few links. Figures 4.3b–4.3k show the clustering results obtained by the MLIRM and several of the related models that were reviewed

[1]http://www.grouplens.org/, as of 2003.

in Section 4.2.3.

As shown in Figure 4.3b, the IRM fails to detect true partitions because it assumes that all observations are relevant to the underlying cluster structure. In contrast, the MLIRM (Figs. 4.3c, 4.3d, and 4.3e) found true partitions by estimating the layer to which each observation was relevant. We also show the solutions obtained by the SIRM and RDIRM, which have a similar assumption, i.e., a background layer blurs cluster structure. In the SIRM (Figs. 4.3f and 4.3g), either the clustering layer or the background layer is selected in an object-wise manner. Therefore, the SIRM cannot consider that an observation is affected by both the cluster structure and object biases. On the other hand, the RDIRM (Figs. 4.3h–4.3k) found more accurate partitions compared to the SIRM. However, the RDIRM considers only one background layer; therefore, only non-links were captured as irrelevant entries.

4.4.2 Real-world Datasets

We applied the MLIRM to three real-world datasets. The first dataset was the "Animal" dataset, which includes relationships between 50 mammals and 85 features. Each feature was rated on a scale of 0–100 for each animal. We prepared binary relational data with a threshold that yields $R_{i,j} = 1$ for all ratings higher than the overall average ratings; i.e., $R_{i,j} = 1(0)$ indicated that animal i has (does not have) feature j. The second dataset was the "Enron" dataset, which contains e-mail transactions among Enron employees. We extracted the e-mail transactions on October 2001, which is when the Enron accounting scandal was first reported. This dataset contains 149 Enron employees. For this dataset, $R_{i,j} = 1(0)$ was used to indicate if an e-mail was (not) sent from employee i to employee j. The last dataset is the "MovieLens" dataset, which contains ratings for 1,682 movies by 943 users on a five-point scale. In our experiment, $R_{i,j} = 1$ when the rating was higher than three points and $R_{i,j} = 0$ otherwise; i.e., $R_{i,j} = 1$ indicates that user i liked movie j.

Table 4.1: Quantitative results for three real-world datasets (i.e., Animal (AN), Enron (EN), and MovieLens (ML)). TE, TL, and $\#KL$ indicate train error (0-1 loss), test data log likelihood, and number of the obtained cluster blocks ($K \times L$), respectively. The best results are highlighted in bold. The parenthesized numbers indicate standard deviations.

AN	IRM	SIRM	RDIRM-prod	RDIRM-sum	MLIRM
TE	0.127 (0.003)	0.129 (0.003)	0.076 (0.005)	0.079 (0.004)	**0.056 (0.008)**
TL	-0.426 (0.016)	-0.480 (0.017)	-0.384 (0.023)	-0.402 (0.022)	**-0.344 (0.021)**
$\#KL$	382.3 (34.7)	376.5 (34.5)	338.9 (41.0)	394.0 (41.2)	**309.0 (26.7)**
EN	IRM	SIRM	RDIRM-prod	RDIRM-sum	MLIRM
TE	0.030 (0.001)	0.028 (0.001)	0.018 (0.001)	0.018 (0.001)	**0.014 (0.002)**
TL	-0.123 (0.006)	-0.133 (0.006)	-0.129 (0.005)	-0.133 (0.007)	**-0.099 (0.005)**
$\#KL$	376.8 (81.8)	391.9 (63.7)	60.3 (15.9)	379.5 (145.2)	**45.7 (12.3)**
ML	IRM	SIRM	RDIRM-prod	RDIRM-sum	MLIRM
TE	0.032 (0.000)	0.032 (0.000)	0.020 (0.000)	0.022 (0.000)	**0.019 (0.000)**
TL	-0.090 (0.000)	-0.093 (0.000)	-0.089 (0.000)	-0.090 (0.000)	**-0.055 (0.000)**
$\#KL$	2091.4 (66.0)	2137.5 (32.4)	**1350.9 (35.0)**	2024.5 (52.7)	**1431.1 (22.7)**

For quantitative comparison, three measurements were used; i.e., train error (0-1 loss), test log likelihood, and number of obtained cluster blocks $K \times L$. Throughout the experiments, we randomly hid 5% of observations during the training period. These hidden entries were used to calculate the test log likelihood, and the remaining entries were used to compute the train error and the number of cluster blocks $K \times L$. The train error ($\in [0, 1]$) indicates the flexibility of an evaluated model. A low train error value means that the model fits better to the training data. The test log likelihood was used to evaluate the predictive robustness of relational models. The test log likelihood is a real-valued measurement indicating the averaged log likelihood of a hidden entry

that takes an actual value; a larger value means the model is more robust in link prediction. The small test log likelihood value indicates that the model overfits the data. In addition, we computed the number of obtained cluster blocks $K \times L$ in order to evaluate the simplicity of the discovered cluster structure. A smaller value means the model abstracts the given data effectively. We calculated each measurement averaged over the last 300 samples of Gibbs iterations.

Table 4.1 lists the computed measures. In the case of every dataset for all measurements, except for the number of cluster blocks $K \times L$ for the "MovieLens" dataset, the MLIRM significantly outperformed the other models. For the "MovieLens" dataset, the MLIRM and RDIRM-prod obtained nearly identical values for the number of $K \times L$ because this dataset was sparse, and a single background layer was sufficient to explain irrelevant entries. As is evident from the table, we have confirmed that the MLIRM performs well in both predictive robustness and its ability to discover simple abstractions from given relational data.

In addition, we qualitatively examined the clustering results for the "Animal" and the "Enron" datasets. To clarify the effects of the proposed model, we compared the results obtained by the IRM and the proposed MLIRM. Figure 4.4 shows the clustering results. As is evident from the figure, we have confirmed that the MLIRM adequately excludes irrelevant entries and finds clear cluster structures. Furthermore, to obtain ideas from learned interaction weights $\boldsymbol{\phi}$, we illustrate the most probable layer obtained for each dataset in Figure 4.6. As can be seen in Fig. 4.6, the proposed MLIRM can flexibly estimate the form of interaction from given data.

Here, we focus closely on the results for the "Animal" dataset. As shown in Figs. 4.4d and 4.4e, there are several features (column objects), in which many related entries are assigned to background layers. To quantify an object's relevance to

each layer u, we compute the relevance scores $S_{1,u}(i)$ and $S_{2,u}(j)$ as follows:

$$S_{1,u}(i) := \frac{\sum_{y \in T_2} \mathbb{I}(r_{i,y} = u)}{J}, S_{2,u}(j) := \frac{\sum_{x \in T_1} \mathbb{I}(r_{x,j} = u)}{I}. \tag{4.85}$$

Table 4.2 lists the top scoring features for each layer. As can be seen in the left hand side of the table, interpretable feature clusters, such as carnivorous features or aquatic features, were obtained. In addition, typical features, such as "meat," "fast," "swims," and "claws" were extracted for each cluster. Furthermore, as is shown in the right hand side of Table 4.2, the MLIRM extracted non-informative features with few links or an extreme number of links. For example, the feature "orange" has links to only four out of 50 animals. Thus, such objects are irrelevant to the underlying cluster structure. Another example is "newworld," which has links to 41 out of 50 animals. Such objects are spam objects; therefore, they are also worthless for clustering. Figure 4.5 depicts the detailed clustering results for the "Animal" dataset. As can be seen from this figure, in the MLIRM, both spamming and passive objects are assigned to their nearest meaningful cluster, which forms at least one dense block. Therefore, clusters obtained by the MLIRM are more informative than those obtained by the IRM. As a result, we have confirmed that the MLIRM has the ability to extract informative clusters and typical objects for each cluster.

Next, we closely look at the results for the "Enron" dataset. As can be seen in Figures 4.4b–4.4h, most e-mails were sent and received by only a few employees. In this case, the IRM forcedly fits itself to the data. As a result, the IRM found many small non-informative clusters (Fig. 4.4b). For the proposed MLIRM, unlike the IRM, nearly all non-links were explained by the background non-link layer, and simple cluster structure are successfully obtained. We confirmed that the clusters obtained by the MLIRM correspond to the major roles of employees, such as "vice presidents," "ordinary employees," "VIPs related to the pipeline business," and "CEOs and presidents." One interesting fact is that only one employee was extracted as a spammer (horizontal dotted line in Fig. 4.4g). This employee was a key person in the

Table 4.2: Top scored features for each layer obtained by the MLIRM for the "Animal" dataset. The table on the left lists examples of obtained feature clusters and the top scoring features for each cluster. The top (bottom) right table lists features highly relevant to the background link (non-link) layer.

feature	$S_{2,\mathrm{fg}}(j)$	$z_{2,j}$
meat	0.98	15
hunter	0.86	15
meatteeth	0.84	15
fierce	0.62	15
fast	1.00	21
tail	0.98	21
agility	0.96	21
swims	0.98	22
coastal	0.98	22
strainteeth	0.98	22
claws	1.00	24
paws	0.84	24
nocturnal	0.76	24

feature	$S_{2,\mathrm{bg1}}(j)$
newworld	0.74
chewteeth	0.72
black	0.60
oldworld	0.57
solitary	0.40
white	0.38

feature	$S_{2,\mathrm{bg0}}(j)$
orange	0.80
yellow	0.80
skimmer	0.78
bush	0.76
stripes	0.76
desert	0.72

Enron scandal and sent e-mails to many other employees in response to the scandal.

These qualitative results indicate that the proposed MLIRM can extract simple and clear cluster structures as well as typical objects within each cluster. Therefore, the proposed MLIRM is useful in discovering interpretable cluster structures from blurred real-world relational data.

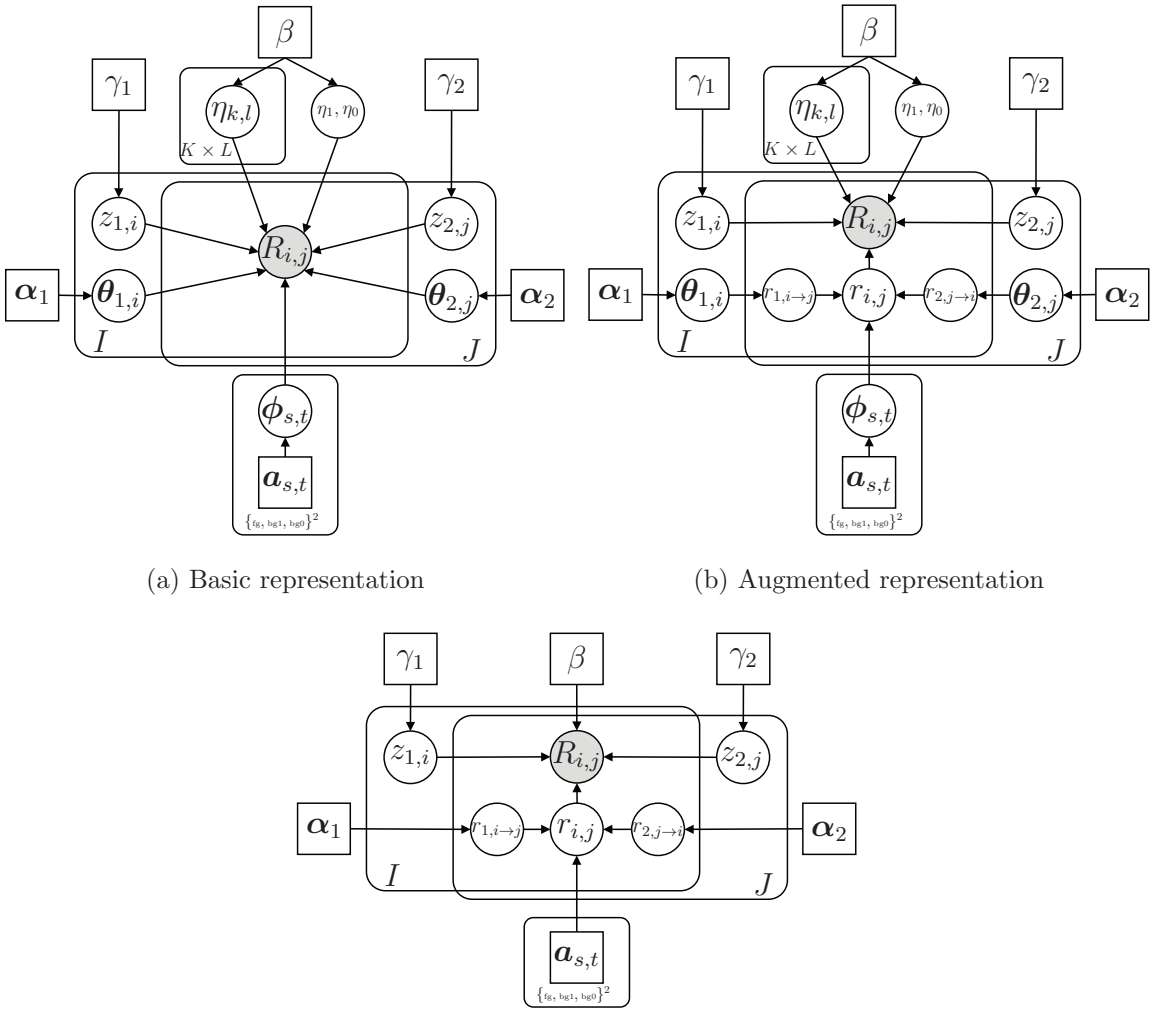
4.5 Chapter Summary

In this chapter, we proposed a new generative framework that extracts a de-blurred cluster structure by estimating the bias of each object and its interactions. In addition, we proposed a new generative model called the MLIRM, which is a concrete instance of the proposed framework that incorporates the IRM. Experiments have confirmed the MLIRM's superiority in predictive accuracy and simplicity of abstraction. Moreover, we observed that the MLIRM successfully found clear bicluster structures and typical objects within each cluster. Therefore, the proposed MLIRM is useful for discovering interpretable cluster structure from blurred real-world relational data.

Finally, we briefly discuss future directions of this chapter. There are two promising aspects for future work.

One is enhancing the computational efficiency. In this study, the posterior inference for the MLIRM and conventional models was performed using a collapsed Gibbs sampling algorithm in order to compere the potential capability of these generative models. However, in general, Gibbs sampling algorithms are relatively computationally expensive compared with more modern algorithms, such as variational inference [70, 32] or gradient-based stochastic methods [11]. Therefore, investigating the scalability and computational efficiency of the MLIRM is an important direction of our future work.

The other is enhancing model capability. The MLIRM includes the IRM for the clustering model; thus, objects are partitioned into non-overlapping clusters. However, mixed or multiple membership assumptions are appropriate in many real-world situations. Therefore, in the future, we plan to apply the proposed multi-layered framework to other advanced clustering models, such as mixed membership [73, 1, 19] or multiple membership models [43, 56, 54]. We are also interested in developing efficient online algorithms for MLIRM in order to accommodate large-scale datasets.

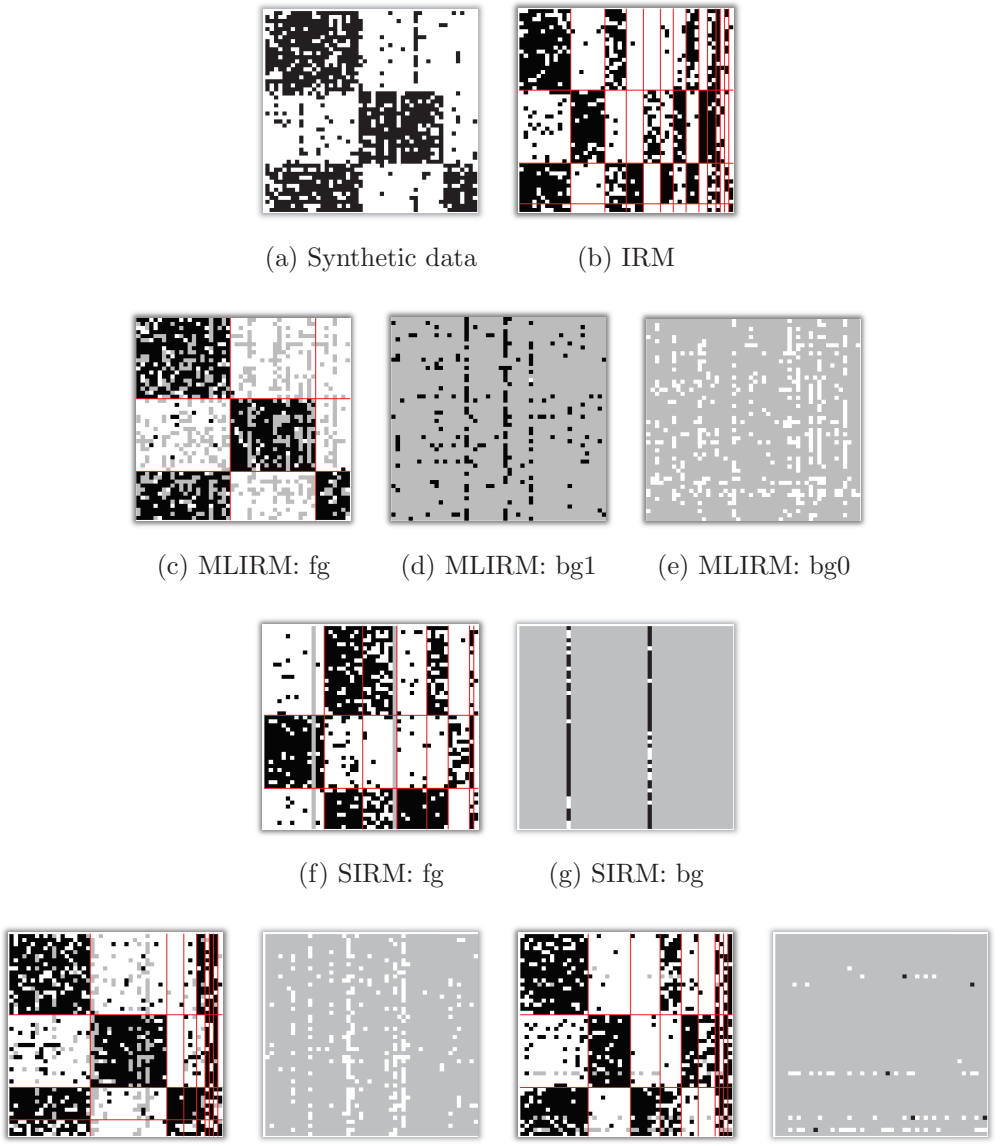


(a) Basic representation

(b) Augmented representation

(c) Marginal representation

Figure 4.2: Graphical representations of the MLIRM. (a) A basic graphical representation of the MLIRM corresponding to Eqs. (4.6)–(4.11) in Section 4.2.2. (b) An augmented representation for the MLIRM with auxiliary variables $r_{1,i\to j}$, $r_{2,j\to i}$, and $r_{i,j}$ (which corresponds to Eqs. (4.15)–(4.22) in Section 4.3.1). (c) The marginal representation obtained by integrating $\boldsymbol{\eta}$, η_1, η_0, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\phi}$ out from the augmented representation (which corresponds to Eq. (4.26) in Section 4.3.1). Note that circle nodes denote random variables, square nodes denote hyperparameters, shaded nodes denote observations, and round-edged squares indicate the number of dimension for individual variables. Directed connections denote probabilistic dependencies.



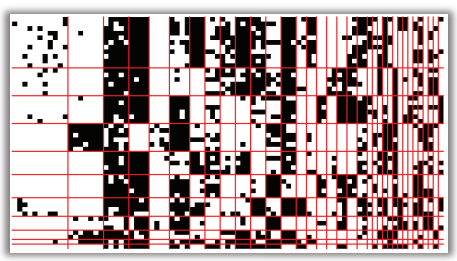
(a) Synthetic data (b) IRM

(c) MLIRM: fg (d) MLIRM: bg1 (e) MLIRM: bg0

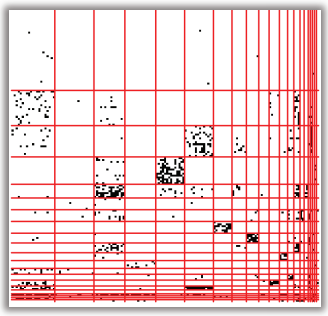
(f) SIRM: fg (g) SIRM: bg

(h) RDIRM-prod: fg (i) RDIRM-prod: bg (j) RDIRM-sum: fg (k) RDIRM-sum: bg

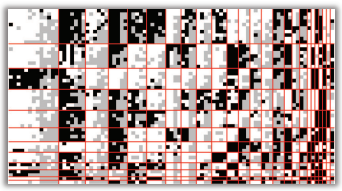
Figure 4.3: Synthetic data example. (a) Synthetic 50×50 relational data (white corresponds to zero, black to one). (b) IRM solution (rows and columns are sorted by the estimated cluster indices). (c)–(e) MLIRM solutions. (c) shows the area assigned to the clustering layer $r_{i,j} = \text{fg}$, (d) to the first background layer $r_{i,j} = \text{bg1}$, and (e) to the background layer $r_{i,j} = \text{bg0}$ (gray area indicates that corresponding entries are assigned to other layers). (f)–(k) Solutions produced by SIRM, RDIRM-prod, and RDIRM-sum, respectively.



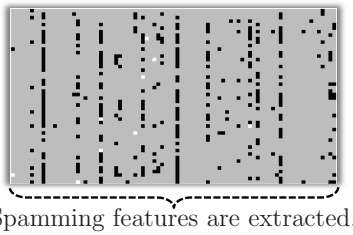
(a) Animal (IRM)



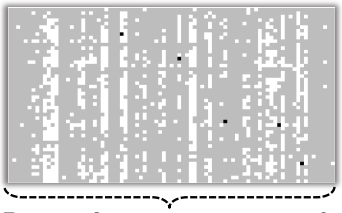
(b) Enron (IRM)



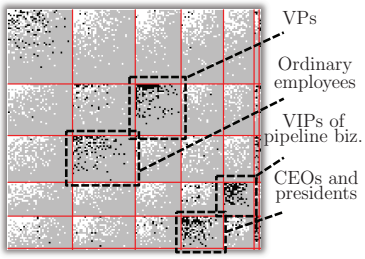
(c) Animal (MLIRM: fg)



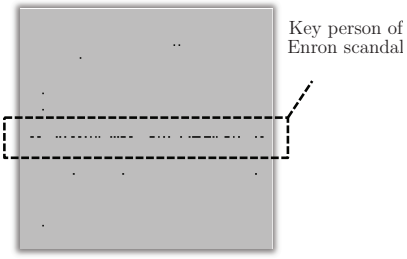
(d) Animal (MLIRM: bg1)



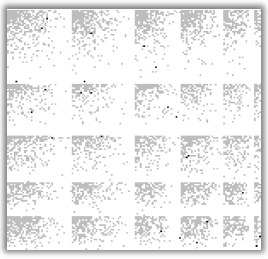
(e) Animal (MLIRM: bg0)



(f) Enron (MLIRM: fg)

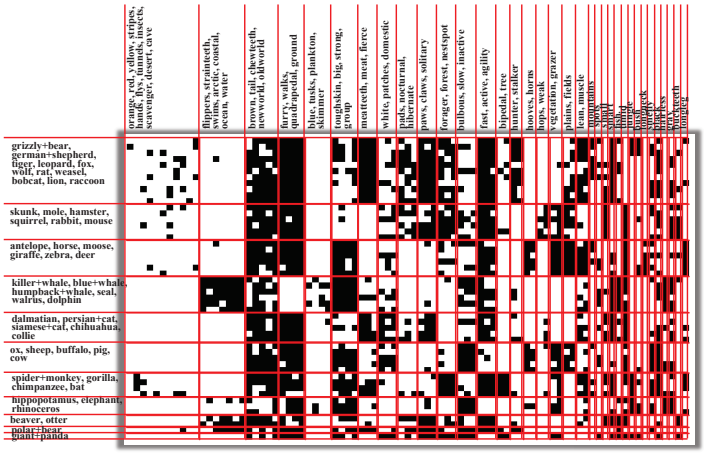


(g) Enron (MLIRM: bg1)



(h) Enron (MLIRM: bg0)

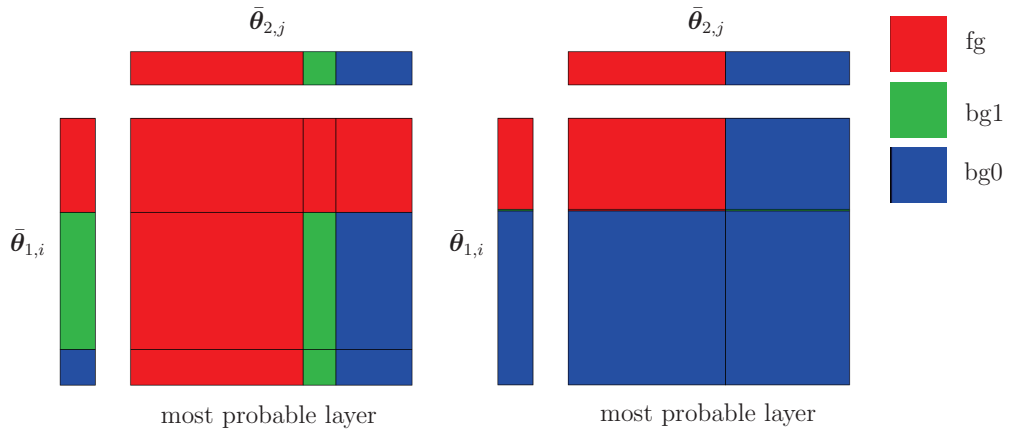
Figure 4.4: Clustering results for the "Animal" and "Enron" datasets. (a) and (b) depict the IRM solutions and (c)–(h) are the MLIRM solutions (objects within each cluster for MLIRM solutions are sorted by descending order of relevance scores $S_{1,\mathrm{fg}}(i)$ and $S_{1,\mathrm{fg}}(j)$).



(a) IRM solution

(b) MLIRM solutions

Figure 4.5: Detailed illustrations of clustering results on "Animal" dataset with object labels. Note that, in the MLIRM solution, two background layers (i.e., bg1 and bg0) are depicted jointly. The gray scaled texts of mammal and feature names for MLIRM solutions indicate the corresponding relevance scores (black to 1.0 and white to 0.0).



(a) Animal (b) Enron

Figure 4.6: (Best viewed in color.) The estimated interaction $\boldsymbol{\phi}$ for (a) "Animal" and (b) "Enron" datasets (color bar of the left hand side (top) of each figure is the mean estimated bias parameters $\bar{\boldsymbol{\theta}}_{1,i}$ ($\bar{\boldsymbol{\theta}}_{2,j}$); the central color map indicates the most probable layer for $r_{i,j}$ conditioned on $r_{1,i\to j}$ and $r_{2,j\to i}$).

Chapter 5

Relevance Modeling with Link Function

In this chapter, in order to develop more computationally efficient relevance-dependent biclustering model, we introduce a link function approach for modeling relevance-dependency. More specifically, we introduce the Relevance-dependent Bernoulli Distribution (R-BD), which is a novel prior distribution for relevance-dependent binary matrices. In our R-BD, a link strength for an entry is defined by three non-negative parameters: a typical link strength common to all entries in the matrix, and two relevance parameters for each row and column objects. Then, an observed link probability is directly calculated by transforming the product of these three non-negative variables into a probability using Bernoulli-Poisson link function [74]. The main advantages of the R-BD is as follows. First, the relevance-modeling in the R-BD do not have to consider any background distributions. Thus, the number of latent variables to be estimated is significantly smaller than those in the RDIRM and MLIRM. Second, the link probability in the R-BD can be modulated widely from 0.0 to 1.0 without introducing complicated mechanism as in the MLIRM. Thus, the effect relevance values in the R-BD is interpretable. Finally, as the all parameters of the R-BD can be completely

marginalized out, we do not have to explicitly estimate R-BD's parameters when performing posterior inference. By incorporating the R-BD as a component distribution, we propose a novel biclustering model termed the Relevance-dependent Infinite Biclustering (R-IB). Thanks to the property of the R-BD, the posterior inference for the R-IB can also be performed using a collapsed Gibbs sampler. Furthermore, the R-IB can be inferred faster than not only the RDIRM and MLIRM, but also the original IRM. Experimental results show that the R-IB extracts more essential bicluster structure with better computational efficiency than conventional models. We further observed that the biclustering results obtained by R-IB facilitate interpretation of the meaning of each cluster.

5.1 Motivations

Few studies consider relevance dependency in biclustering. The Relevance Dependent Infinite Relational Model (RDIRM) [48, 49] assumes that only a subset of entries is relevant to the cluster structure.

More specifically, in the RDIRM, mixing parameters $\rho_{1,i}, \rho_{2,j} \in [0, 1]$ are introduced for each object, and observations are drawn from a mixture of foreground $\eta_{z_{1,i},z_{2,j}}$ and background η_0 densities as follows:

$$\begin{aligned} r_{1,i\to j} \,|\, \rho_{1,i} &\sim \text{Bernoulli}(\rho_{1,i}), \\ r_{2,j\to i} \,|\, \rho_{2,j} &\sim \text{Bernoulli}(\rho_{2,j}), \\ r_{i,j} &= f(r_{1,i\to j}, r_{2,j\to i}), \\ R_{i,j} \,|\, z_{1,i}, z_{j,2}, \boldsymbol{\eta}, \eta_0, r_{i,j} &\sim \text{Bernoulli}(r_{i,j} \times \eta_{z_{1,i},z_{2,j}} + (1 - r_{i,j}) \times \eta_0), \end{aligned} \tag{5.1}$$

where $f(\cdot, \cdot)$ is a Boolean function typically set to the logical product.

Although the aim of the RDIRM is to exclude non-informative entries as noise, this mechanism can be considered an approach for modeling relevance dependency because

mixing parameters $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ affect the observed link probabilities regardless of a given object's cluster membership.

However, there is a number of drawbacks in their approach. First, in the RDIRM, a link probability for an entry depends on many internal parameters: foreground probability $\eta_{z_{1,i},z_{2,j}}$, background probability η_0, mixing parameters $\rho_{1,i}, \rho_{2,j}$, and Boolean function $f(\cdot,\cdot)$. This makes the effect of the relevance on link probabilities too complex to interpret. Second, to infer the RDIRM, not only $I+J$ cluster assignments $\boldsymbol{z}_1, \boldsymbol{z}_2$ must be estimated but also $I \times J$ latent variables $\boldsymbol{r}_1, \boldsymbol{r}_2$. Thus, the RDIRM can be applied to only very small relational data. Finally, no reasonable strategy is available to select the Boolean function $f(\cdot,\cdot)$. Ohama *et al.* [50] tackled this problem by assuming the prior for Boolean functions. However, the interpretability of relevance is degraded as the estimated probabilistic Boolean function becomes increasingly complex.

In this chapter, we introduce the Relevance-Dependent Bernoulli Distribution (R-BD) as a prior for relevance-dependent binary matrices. In the R-BD, instead of the mixed-membership modeling in the RDIRM and MLIRM, a Bernoulli-Poisson link function [74] is used for relevance-dependency modeling. Therefore, the posterior inference for the R-BD can be performed efficiently without partitioning given relational data into foreground and background part. By incorporating the R-BD as a observation model of the IRM, we propose the novel Relevance-Dependent Infinite Biclustering (R-IB) model, which automatically estimates the number of clusters. Posterior inference for the R-IB can be performed efficiently using a collapsed Gibbs sampler because the parameters of the R-IB model can be fully marginalized out.

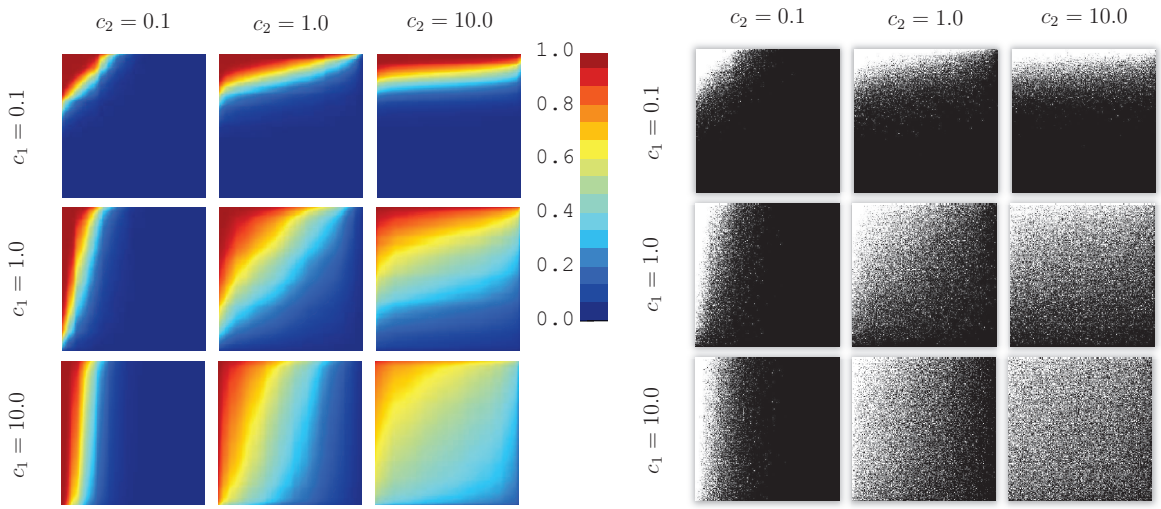


Figure 5.1: Outcomes drawn from the R-BD ($\lambda = 0.6931$) with different Dirichlet parameters. (a) shows probability densities and (b) shows binary matrices drawn from corresponding density, where black corresponds to 0 and white to 1.

5.2 Relevance-dependent Infinite Biclustering (R-IB)

First, we introduce the R-BD: a novel prior for relevance-dependent binary matrices. Then, by incorporating the R-BD, we propose the R-IB model.

5.2.1 Relevance-Dependent Bernoulli Distribution

To design a prior distribution for an $I \times J$ relevance-dependent binary matrix $\boldsymbol{x}$, we consider three non-negative parameters λ, $\psi_{1,i}$, and $\psi_{j,2}$. The first parameter λ in the range $[0, +\infty)$ is a typical link strength that controls the overall density for matrix $\boldsymbol{x}$. The remaining parameters $\psi_{1,i}$ and $\psi_{2,j}$ (also in $[0, +\infty)$) are the relevance parameters for the i-th row and the j-th column, respectively. Then, we define the relevance-

dependent link strength for an entry $x_{i,j}$ by multiplying these parameters as $\psi_{i,1}\psi_{j,2}\lambda$. Finally, to obtain a binary random variable, we define Relevance-dependent Bernoulli distribution (R-BD) as follows:

$$x_{i,j} \,|\, \psi_{1,i}, \psi_{2,j}, \lambda \sim \text{Bernoulli}\left(1 - e^{-\psi_{1,i}\psi_{2,j}\lambda}\right), \tag{5.2}$$

where the function $1 - e^{-\cdot}$ is the Bernoulli-Poisson (BerPo) link function [74] that transform a non-negative variable s into a probability.

The relevance modeling in our R-BD is more interpretable than that in the RDIRM, because the effect of relevance is defined by a simple multiplication of non-negative variables.

Another remarkable property of R-BD is that all internal parameters (i.e., λ, $\boldsymbol{\psi_1}$, and $\boldsymbol{\psi_2}$) can be marginalized out. Following the property of BerPo link, Eq. (5.2) can be equivalently rewritten by truncating a Poisson random variable $x^*_{i,j}$ as

$$\begin{aligned} x^*_{i,j} \,|\, \psi_{1,i}, \psi_{2,j}, \lambda &\sim \text{Poisson}(\psi_{1,i}\psi_{2,j}\lambda), \\ x_{i,j} &= \mathbb{I}(x^*_{i,j} \geq 1), \end{aligned} \tag{5.3}$$

where $\mathbb{I}(\cdot)$ is 1 if the predicate holds and is 0 otherwise. Posterior sampling of $x^*_{i,j}$ can be easily performed as follows:

$$x^*_{i,j} \,|\, x_{i,j}, \psi_{1,i}, \psi_{2,j}, \lambda \sim \begin{cases} \delta(0), & \text{if } x_{i,j} = 0 \\ \text{ZTP}(\psi_{1,i}\psi_{2,j}\lambda). & \text{if } x_{i,j} = 1 \end{cases} \tag{5.4}$$

Note that $\delta(0)$ is a point mass at zero and $\text{ZTP}(\cdot)$ denotes a *zero-truncated Poisson* distribution [21], which is also known as the conditional Poisson distribution [10] or the positive Poisson distribution [65]. This representation enables the construction of conjugate priors for R-BD parameters. Assuming gamma and Dirichlet priors as $\lambda \sim \text{Gamma}(a, b)$ [1], $\{\psi_{1,i}\}_{i=1}^{I}/I \sim \text{Dirichlet}(c_1)$, and $\{\psi_{2,j}\}_{j=1}^{J}/J \sim \text{Dirichlet}(c_2)$, we

[1] $\text{Gamma}(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b, i.e., $P(\lambda \,|\, a, b) = \lambda^{a-1}e^{-b\lambda}/G(a, b)$ where $G(a, b) = \Gamma(a)/b^a$. $\Gamma(\cdot)$ denotes the gamma function.

obtain a closed-form marginal likelihood for auxiliary counts $\boldsymbol{x}^*$ as follows:

$$P(\boldsymbol{x}^*) = \frac{1}{\prod_{i=1}^{I}\prod_{j=1}^{J} x^*_{i,j}!} \times \prod_{i=1}^{I} \frac{\Gamma(c_1 + M_{i,\cdot})}{\Gamma(c_1)} \times \prod_{j=1}^{J} \frac{\Gamma(c_2 + M_{\cdot,j})}{\Gamma(c_2)} \times \frac{I^M \Gamma(Ic_1)}{\Gamma(Ic_1 + M)} \times \frac{J^M \Gamma(Jc_2)}{\Gamma(Jc_2 + M)} \times \frac{G(a + M, b + IJ)}{G(a, b)}, \tag{5.5}$$

where $M_{i,\cdot} = \sum_{j=1}^{J} x^*_{i,j}$, $M_{\cdot,j} = \sum_{i=1}^{I} x^*_{i,j}$, and $M = \sum_{i=1}^{I}\sum_{j=1}^{J} x^*_{i,j}$. Thus, the parameters for R-BD no longer need to be estimated explicitly because they have been marginalized out. This gives the R-BD an affinity with collapsed inference.

Figure 5.1 depicts the random binary matrices drawn from R-BD with different Dirichlet parameters. Although the binary matrices drawn from R-BD indicate various density patterns, the expected link strengths for these matrices are equal to exactly λ. Therefore, the estimated value of λ can be interpreted as a representative link strength value of a given binary matrix.

5.2.2 Relevance-Dependent Infinite Biclustering

Here, we describe the proposed R-IB model. CRP(γ) denotes a CRP with concentration parameter γ. The full description of the R-IB model, incorporating R-BD to the observation model of the IRM, is as follows:

$$\begin{aligned}
z_{1,i} \,|\, \gamma_1 &\sim \mathrm{CRP}(\gamma_1), \\
z_{2,j} \,|\, \gamma_2 &\sim \mathrm{CRP}(\gamma_2), \\
\boldsymbol{\psi}_{1,k}/m_{1,k} \,|\, c_1, \boldsymbol{z}_1 &\sim \mathrm{Dirichlet}(\overbrace{c_1, \ldots, c_1}^{m_{1,k}}), \\
\boldsymbol{\psi}_{2,l}/m_{2,l} \,|\, c_2, \boldsymbol{z}_2 &\sim \mathrm{Dirichlet}(\overbrace{c_2, \ldots, c_2}^{m_{2,l}}), \\
\lambda_{k,l} \,|\, a, b &\sim \mathrm{Gamma}(a, b), \\
R_{i,j} \,|\, z_{1,i}, z_{2,j}, \psi_{1,i}, \psi_{2,j}, \boldsymbol{\lambda} &\sim \mathrm{Bernoulli}\left(1 - e^{-\psi_{1,i}\psi_{2,j}\lambda_{z_{1,i},z_{2,j}}}\right),
\end{aligned} \tag{5.6}$$

where $m_{1,k}$ ($m_{2,l}$) is the number of row (column) objects assigned to cluster k (l). Note that $\boldsymbol{\psi}_{1,k}$ ($\boldsymbol{\psi}_{2,l}$) is a set of relevance parameters $\psi_{1,i}$ ($\psi_{2,j}$), where $z_{1,i} = k$ ($z_{2,j} = l$).

Thanks to the conjugacy between R-BD and its priors, by introducing auxiliary Poisson counts $\boldsymbol{R}^*$, the model parameters $\boldsymbol{\lambda}$, $\boldsymbol{\psi}_1$, and $\boldsymbol{\psi}_2$ can be marginalized out. The marginal likelihood for $\boldsymbol{R}^*$, given $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, is then given as

$$\begin{aligned} P(\boldsymbol{R}^* \mid \boldsymbol{z}_1, \boldsymbol{z}_2) =& \frac{1}{\prod_{i=1}^{I}\prod_{j=1}^{J} R^*_{i,j}!} \times \prod_{i=1}^{I} \frac{\Gamma(c_1 + M_{i,\cdot})}{\Gamma(c_1)} \times \prod_{j=1}^{J} \frac{\Gamma(c_2 + M_{\cdot,j})}{\Gamma(c_2)} \\ & \times \prod_{k=1}^{K} \frac{m_{1,k}^{M_{k,\cdot}} \Gamma(m_{1,k} c_1)}{\Gamma(m_{1,k} c_1 + M_{k,\cdot})} \times \prod_{l=1}^{L} \frac{m_{2,l}^{M_{\cdot,l}} \Gamma(m_{2,l} c_2)}{\Gamma(m_{2,l} c_2 + M_{\cdot,l})} \\ & \times \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{G(a + M_{k,l}, b + m_{1,k} m_{2,l})}{G(a,b)}, \end{aligned} \tag{5.7}$$

where $M_{k,l} = \sum_{i=1}^{I}\sum_{j=1}^{J} R^*_{i,j} \mathbb{I}(z_{1,i} = k)\mathbb{I}(z_{2,j} = l)$, $M_{k,\cdot} = \sum_{l=1}^{L} M_{k,l}$, and $M_{\cdot,l} = \sum_{k=1}^{K} M_{k,l}$.

Figure 5.2 shows an R-IB solution for a synthetic dataset, in which link probabilities are distorted by object relevance. As can be seen in Fig. 5.2a, a 3×3 bicluster structure is present in the data. As shown in Fig. 5.2b, the IRM fails to extract the true partitions, because the IRM assumes uniform density within each block. In contrast, the R-IB (Fig. 5.2c) successfully finds the true partitions by estimating relevance values.

5.3 Inference

Posterior inference for the R-IB can be performed via collapsed Gibbs sampling. As the parameters of R-BD have been marginalized out, the only variables we have to estimate are cluster assignments $\boldsymbol{z}_1, \boldsymbol{z}_2$ and auxiliary counts $\boldsymbol{R}^*$.

As $z_{1,i}$ and $z_{2,j}$ can be sampled in the same way, we concentrate on $z_{1,i}$. Using (5.7) and the likelihood for the CRP, given $\boldsymbol{R}^*$, the posterior probability that the i-th object

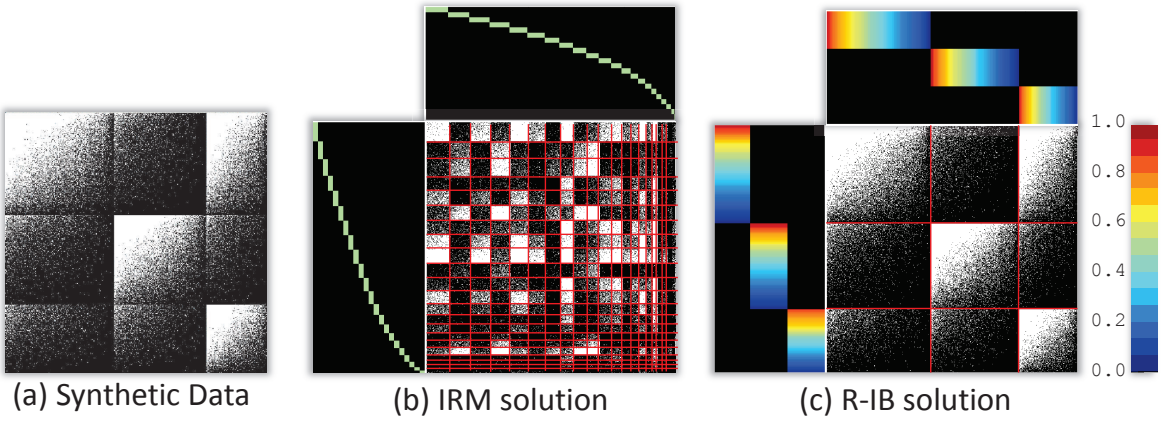


Figure 5.2: Synthetic example: (a) 500×500 relational data; (b) IRM solution; (c) R-IB solution. In (b) and (c), the left and top matrices indicate $\boldsymbol{z}_1$ and $\boldsymbol{z}_2^\top$ in a 1-of-K representation, respectively. Colored areas in $\boldsymbol{z}_1$ and $\boldsymbol{z}_2^\top$ indicate relevance parameters for corresponding objects. For an intuitive understanding, each relevance parameter ψ is transformed into a probability in $[0, 1]$ as $1 - e^{-\log(2) \times \psi}$.

is assigned to cluster k^* is given by

$$
\begin{aligned}
&P(z_{1,i} = k^* \,|\, -) \\
&\propto \begin{cases}
m_{1,k^*}^{-i} \times \dfrac{\left(m_{1,k^*}^{+i}\right)^{M_{k^*,\cdot}^{+i}} \Gamma(m_{1,k^*}^{+i} c_1) \Gamma(m_{1,k^*}^{-i} c_1 + M_{k^*,\cdot}^{-i})}{\left(m_{1,k^*}^{-i}\right)^{M_{k^*,\cdot}^{-i}} \Gamma(m_{1,k^*}^{-i} c_1) \Gamma(m_{1,k^*}^{+i} c_1 + M_{k^*,\cdot}^{+i})} \\
\qquad \times \displaystyle\prod_{l=1}^{L} \frac{G(a + M_{k^*,l}^{+i}, b + m_{1,k^*}^{+i} m_{2,l})}{G(a + M_{k^*,l}^{-i}, b + m_{1,k^*}^{-i} m_{2,l})}, & \text{if } m_{1,k^*}^{-i} > 0 \\
\gamma_1 \times \dfrac{\Gamma(c_1)}{\Gamma(c_1 + M_{i,\cdot})} \\
\qquad \times \displaystyle\prod_{l=1}^{L} \frac{G(a + M_{k^*,l}^{+i}, b + m_{1,k^*}^{+i} m_{2,l})}{G(a, b)}, & \text{if } m_{1,k^*}^{-i} = 0
\end{cases}
\end{aligned}
\tag{5.8}
$$

where superscript $-i$ indicates that the corresponding statistic is computed while excluding the i-th row object. Conversely, $+i$ means that the corresponding statistic is computed while including the i-th row object in cluster k^*.

From (5.4), the posterior sampling for $R^*_{i,j}$ is given by

$$R^*_{i,j} \,|\, - \sim \begin{cases} \delta(0), & \text{if } R_{i,j} = 0 \\ \mathrm{ZTP}(\psi_{1,i}\psi_{2,j}\lambda_{z_{1,i},z_{2,j}}). & \text{if } R_{i,j} = 1 \end{cases} \tag{5.9}$$

Note that explicit samples for $\boldsymbol{\lambda}$, $\boldsymbol{\psi}_1$, and $\boldsymbol{\psi}_2$ are only required during the sampling of $\boldsymbol{R}^*$, and are drawn as follows:

$$\lambda_{k,l} \,|\, - \sim \mathrm{Gamma}(a + M_{k,l}, b + m_{1,k} \times m_{2,l}), \tag{5.10}$$

$$\boldsymbol{\psi}_{1,k}/m_{1,k} \,|\, - \sim \mathrm{Dirichlet}(c_1 + \boldsymbol{M}_{1,k}), \tag{5.11}$$

$$\boldsymbol{\psi}_{2,l}/m_{2,l} \,|\, - \sim \mathrm{Dirichlet}(c_2 + \boldsymbol{M}_{2,k}), \tag{5.12}$$

where $c_1 + \boldsymbol{M}_{1,k}$ ($c_2 + \boldsymbol{M}_{2,l}$) is the set of $c_1 + M_{i,\cdot}$ ($c_2 + M_{\cdot,j}$) in row (column) cluster k (l).

As the sampling for $\boldsymbol{R}^*$ is computationally insignificant compared with that for $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, both the R-IB and IRM require the $O((I+J)KL)$ computation for each iteration. However, the computation of a beta function required for the IRM is more expensive than that of a gamma function required for the R-IB. As a result, computational time of the R-IB is significantly shorter than that of both the RDIRM and IRM.

Estimating Hyperparameters

The hyperparameters for the R-IB (i.e., γ_1, γ_2, c_1, c_2, a, and b) can also be sampled assuming the gamma prior $\mathrm{Gamma}(e_0, f_0)$. Because of the conjugacy between gamma distributions, the rate parameter b is straightforwardly updated as $b \sim \mathrm{Gamma}(e_0 + aKL, f_0 + \sum_{k,l} \lambda_{k,l})$. For the remaining hyperparameters, posterior sampling is performed using the data augmentation [16, 45, 69, 74] technique that we have described in Sec. 4.3.3.

For the shape parameter a, by applying Eq. (4.61) to each term of the $K \times L$ product in (5.7), we obtain a joint distribution over a and $q_{k,l}$. Then, assuming $a \sim$

Gamma(e_0, f_0), a can be updated as

$$q_{k,l} \,|\, - \sim \text{Antoniak}(M_{k,l}, a), \tag{5.13}$$

$$a \,|\, - \sim \text{Gamma}\left(e_0 + \sum_{k=1}^{K}\sum_{l=1}^{L} q_{k,l}, f_0 + \sum_{k=1}^{K}\sum_{l=1}^{L} \ln\frac{b + m_{1,k}m_{2,l}}{b}\right), \tag{5.14}$$

where Antoniak$(M_{k,l}, a)$ is an Antoniak distribution [3]. This is the distribution of the number of occupied tables if $M_{k,l}$ customers are assigned to one of an infinite number of tables with CRP(a), and is sampled as $q_{k,l} \sim \sum_{p=1}^{M_{k,l}} \text{Bernoulli}(a/(a+p-1))$.

Similarly, by applying Eq. (4.60) and (4.61) to the terms related to c_1 in Eq. (5.7), we obtain a joint distribution over c_1, $p_{1,k}$, and $q_{1,i}$. Consequently, c_1 is updated as

$$q_{1,i} \,|\, - \sim \text{Antoniak}(M_{i,\cdot}, c_1), \tag{5.15}$$

$$p_{1,k} \,|\, - \sim \text{Beta}(m_{1,k}c_1, M_{k,\cdot}), \tag{5.16}$$

$$c_1 \,|\, - \sim \text{Gamma}\left(e_0 + \sum_{i=1}^{I} q_{1,i}, f_0 - \sum_{k=1}^{K} m_{1,k} \ln p_{1,k}.\right). \tag{5.17}$$

Note that c_2 can be sampled in the same way as c_1.

Furthermore, assuming a gamma prior for γ_1, the posterior is proportional to $\gamma_1^{e_0+K-1}e^{-f_0\gamma_1}\Gamma(\gamma_1)/\Gamma(I+\gamma_1)$. We then update γ_1 as

$$p \,|\, - \sim \text{Beta}(\gamma_1, I), \tag{5.18}$$

$$\gamma_1 \,|\, - \sim \text{Gamma}(e_0 + K, f_0 - \ln p). \tag{5.19}$$

Note that γ_2 can be sampled in the same way as γ_1.

5.4 Experiments

We present experimental results obtained using real-world datasets. The purposes of the experiments are as follows:

- To quantitatively show that the R-IB can capture more essential cluster structures with better computational efficiency than the IRM and RDIRM (Sec. 5.4.2).

- To show the usefulness of relevance-dependent biclustering results obtained by the R-IB in understanding the meaning of each cluster (Sec. 5.4.3).

In all the experiments, we also fit all hyperparameters of both the proposed and baseline models assuming the same gamma priors (Gamma(1.0,1.0)).

5.4.1 Datasets

The first dataset was the Animal [55] dataset, which maps relationships between 50 mammals and 85 attributes. Each attribute is rated on a scale of 0–100 for each animal. We prepared binary relational data with a threshold that yielded $R_{i,j} = 1$ for all ratings higher than the overall average rates. Therefore, $R_{i,j} = 1(0)$ indicated that the i-th animal had (or lacked) the j-th attribute. The second dataset was the Enron [35] dataset, which comprises e-mails sent between Enron employees. We extracted e-mail transactions between August and October 2001, and constructed three relational datasets: Enron08, Enron09, and Enron10. These contained e-mail transactions between 149 employees in the corresponding month. For these datasets, $R_{i,j} = 1(0)$ was used to indicate whether an e-mail was, or was not, sent by the i-th employee to the j-th employee. The final dataset was the MovieLens[2] dataset, which comprises five-point scale ratings of 1,682 movies submitted by 943 users. For this dataset, we set $R_{i,j} = 1$ when the rating was higher than three and $R_{i,j} = 0$ otherwise, so that $R_{i,j} = 1(0)$ indicated whether or not the i-th user liked the j-th movie. The densities of the Animal, Enron08, Enron09, Enron10, and MovieLens datasets were 0.368, 0.015, 0.016, 0.026, and 0.035, respectively (summarized in Table 5.1).

Table 5.1: Summary of the datasets.

Datasets	I	J	Density
Animal	50	85	0.368
Enron08	179	179	0.015
Enron09	179	179	0.016
Enron10	179	179	0.026
MovieLens	943	1,682	0.035

5.4.2 Quantitative Comparison

Many real-world relational data contains many zero entries. Thus, in order to evaluate the ability of the R-IB to capture essential bicluster structure, we evaluated the link prediction ability for held-out entries by calculating the averaged Area Under the Curve of both the Precision-Recall curve (AUC-PR) and the ROC curve (AUC-ROC) [12]. We compared three biclustering models: the IRM, the RDIRM, and the R-IB. We ran 4,000 Gibbs iterations for each model on each dataset and used the final 500 iterations to calculate the measurement. All scores were calculated using 10-fold cross validation, and the overall average and deviation were reported.

Table 5.2 lists the results. As can be seen, R-IB significantly outperformed the RDIRM with almost all datasets. For the AUC-PR, the IRM demonstrated the best performance for only the Animal dataset. Compared with the IRM, the other models require additional parameters to be estimated. This caused the RDIRM, and R-IB to overfit the data, because the Animal dataset was too small to allow the underlying cluster structure to be generalized. However, the difficulty in obtaining insights from the data increased as the datasets became larger. As we consider the performance with larger datasets to be a more important criterion, the results demonstrated the

[2]http://www.grouplens.org/, as of 2003.

Table 5.2: Computed AUC-PR (top) and AUC-ROC (bottom) on real-world datasets. Best results are highlighted in bold. Parenthesized numbers indicate standard deviations.

AUC-PR	IRM	RDIRM	R-IB
Animal	**0.811** (0.036)	0.752 (0.053)	0.802 (0.026)
Enron08	0.274 (0.069)	0.204 (0.048)	**0.289** (0.074)
Enron09	0.271 (0.049)	0.213 (0.045)	**0.296** (0.055)
Enron10	0.352 (0.040)	0.310 (0.033)	**0.381** (0.042)
MovieLens	0.410 (0.006)	0.413 (0.006)	**0.447** (0.006)
AUC-ROC	IRM	RDIRM	R-IB
Animal	**0.886** (0.016)	0.846 (0.026)	0.878 (0.017)
Enron08	0.880 (0.024)	0.859 (0.036)	**0.884** (0.026)
Enron09	0.883 (0.026)	0.879 (0.021)	**0.892** (0.026)
Enron10	0.896 (0.019)	0.893 (0.011)	**0.905** (0.018)
MovieLens	0.934 (0.001)	0.937 (0.001)	**0.940** (0.001)

superiority of our R-IB over conventional models in link prediction accuracy.

We also evaluated the Test Data Log-Likelihood (TDLL), which is one of the most popular measures for evaluating generalization ability of statistical models. Table 5.3 lists the results. As can be seen, the RDIRM indicated better performances than those of the proposed R-IB model. In our R-IB, the binary observations are modeled via asymmetric BerPo link function. Therefore, the R-IB tends to fit strongly to non-zero observations. Consequently, in terms of the likelihood, the RDIRM indicated better performances. However, as we have already discussed in this section, real-world binary relational data is often very sparse. Therefore, the AUC is more important measure than the TDLL for evaluating the usefulness of the model in many real-world situations.

Finally, the average number of Gibbs iterations within 5 min was used as the metric

Table 5.3: Computed TDLL on real-world datasets. Best results are highlighted in bold. Parenthesized numbers indicate standard deviations.

TDLL	IRM	RDIRM	R-IB
Animal	**-0.415** (0.031)	-0.467 (0.026)	-0.499 (0.044)
Enron08	-0.061 (0.009)	**-0.060** (0.007)	-0.062 (0.010)
Enron09	-0.067 (0.011)	**-0.062** (0.006)	-0.065 (0.011)
Enron10	-0.085 (0.013)	**-0.081** (0.006)	-0.085 (0.012)
MovieLens	-0.092 (0.001)	-0.091 (0.001)	**-0.089 (0.001)**

to evaluate the computational efficiency of the different models. As shown in Fig. 5.3, our R-IB overwhelmingly outperformed the RDIRM. Even in the worst case, posterior sampling for the R-IB was 16.1 times faster than that for the RDIRM. Furthermore, the proposed R-IB significantly outperformed the baseline standard biclustering model (i.e., the IRM), providing experimental confirmation of the computational efficiency of our model, as discussed in Sec. 5.3.

These quantitative results confirmed that the R-IB can extract more essential bicluster structures with better computational efficiency than conventional models.

Discussion

Although the quantitative results described in Sec. 5.4.2 illustrate the computational efficiency of the proposed R-IB model, further evaluations of the scalability of the model are required. Here, we briefly discuss the computational properties of the R-IB model.

As we have shown in Sec. 5.3, the R-IB requires an $O((I+J)KL)$ computation for each Gibbs iteration. This is significantly faster than the RDIRM, which requires an $O(IJ)$ computation to update latent variables $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. However, in general, size of latent blocks KL underlying relational data increase as the size of given data grows.

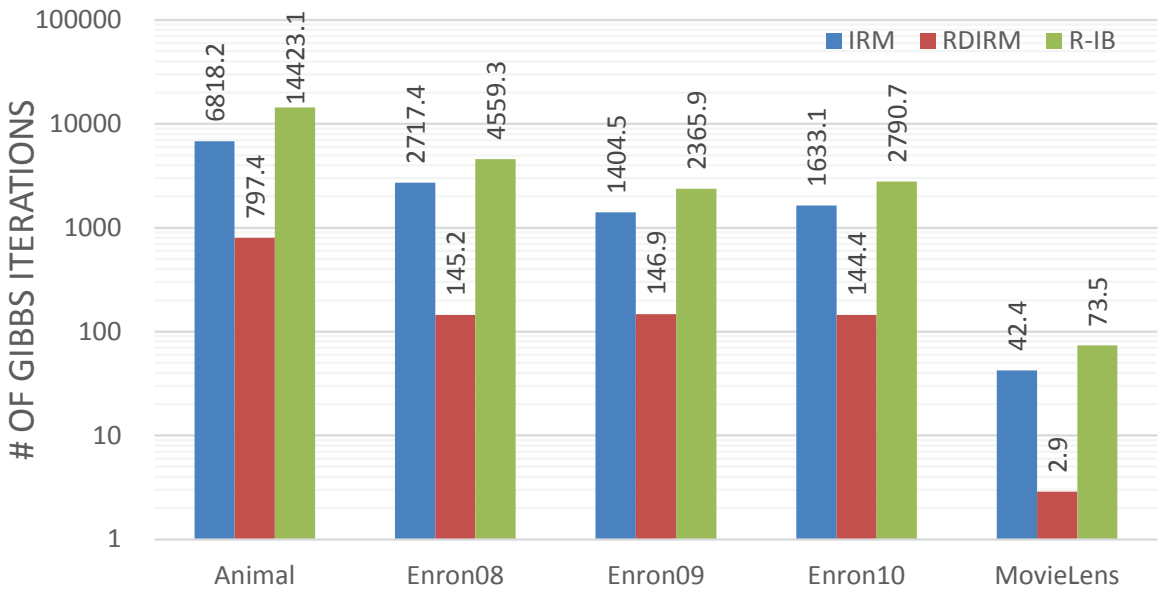


Figure 5.3: Average number of Gibbs iterations per five minutes in logarithmic scale. All the models were implemented in Java and run on a PC with an Intel® Xeon® 2.7 GHz CPU.

Therefore, it is worth exploring more computationally efficient inference algorithms for the R-IB in order to apply the model to huge relational data (e.g., MovieLens20M, which is the largest MovieLens dataset).

Scalability of the R-IB depends on the choice of the inference algorithm. In this study, to evaluate the potential capability of the proposed and conventional models, we used collapsed Gibbs samplers, which enable us to infer the models without any approximations. However, in terms of computational efficiency, collapsed Gibbs sampler is relatively expensive because it cannot be parallelized straightforwardly. Thus, to evaluate the applicability for large scale data, we need to develop more advanced inference algorithms (e.g., variational inference and stochastic optimization) with affinities for parallelized or distributed computation. Such advanced algorithms will be investigated as part of our future work.

5.4.3 Qualitative Comparison

We qualitatively compared our outcomes for the Animal and Enron09 datasets with those obtained using the IRM.

Figure 5.4 shows the clustering results on the Animal dataset. As can be seen from Fig. 5.4a and b, our R-IB abstracted relational data into relevance-dependent blocks in such a way that each block followed the R-BD, whereas blocks obtained using the IRM followed a uniform density. Thus, the IRM form non-informative clusters for irrelevant objects with few links (e.g., column cluster 1 in Fig. 5.4a). In contrast, as Fig. 5.4b shows, all clusters obtained by the proposed R-IB were informative because they were related to at least one meaningful (dense) block.

To assess the contents of the extracted clusters, Fig. 5.5 shows the content of several clusters obtained by R-IB. In each cluster, we can understand its meaning by inspecting only a few top-ranked objects. For example, "Meatteeth" and "Fierce" in column cluster 3 clearly suggest that the cluster denotes carnivorous features. Similarly, other top-ranked objects (e.g., "(Eat) fish," "Quadrapedal," and "Vegetation") also facilitate interpretation of the corresponding clusters. Objects with smaller relevance values were also interesting. For example, "Meatteeth" and "Stalker" in column cluster 3 are definitely related because many carnivorous mammals stalk other animals to prey on them. However, as the third column of column cluster 3 (Fig. 5.5) shows, the IRM assigned them to different clusters because the IRM does not consider the heterogeneity of objects' relevance.

The results obtained by R-IB on the Animal dataset further suggested that the relevance of column objects varied more widely than that of row objects (see the relevance values listed in Fig. 5.5). Here all row objects were selected from a specified category (i.e., mammals). Thus, the row objects followed the underlying cluster structure with the same degree of clarity. In contrast, the attributes of the column objects covered a range of categories such as habitat, favorite food, appearance, and behavioral char-

acteristics. Therefore, the relevance of the column objects was heterogeneous. Thus, the R-IB was shown to be able not only to extract the relevance-dependent bicluster structure but also to assess the necessity of relevance modeling of an arbitrary dataset.

Figure 5.6 shows the solutions obtained from the Enron09 dataset. As can be seen, the IRM produced several non-informative large blocks containing objects with few links (e.g., block A in Fig. 5.6a). Although the IRM also produced a comparatively large cluster block comprising many moderately strongly linked objects (block B in Fig. 5.6a), it was difficult to understand the meaning of the block. In contrast, R-IB produced a more interpretable bicluster structure (Fig. 5.6b). In the R-IB solution, the relevance parameters for objects with few links yielded small values, and these objects were assigned to the nearest meaningful cluster. Therefore, almost all the clusters produced by the R-IB were informative and worthy of inspection. In block C of Fig. 5.6b, the top five relevant row objects were four vice presidents and an anonymized person, and the top two relevant column objects were presidents. This allowed us to assume that the main role of block C could be understood as "reports from vice presidents to presidents." Similarly, blocks D and E in Fig. 5.6b were interpreted as "mails between cash analysts" and "reports from employees to the chief financial officer," respectively.

These results confirmed that the R-IB successfully extracted a relevance-dependent bicluster structure, allowing deep insights to be gained from real-world relational data.

5.5 Chapter Summary

In this chapter, we proposed the R-BD as a prior distribution for relevance-dependent binary matrices. We further proposed conjugate priors for the R-BD to make collapsed inferences available. By incorporating R-BD as an observation model, we introduced a novel infinite biclustering model (i.e., R-IB) that is able to extract a relevance-

dependent bicluster structure from relational data with an unknown number of clusters. Finally, we proposed an efficient collapsed Gibbs sampler to infer the R-IB. Experiments using real-world datasets confirmed that the R-IB was able to extract more essential clusters with better computational efficiency than conventional models. We further confirmed that relevance-dependent clusters obtained by the R-IB were more interpretable than those obtained by standard biclustering.

There are two promising directions for future work. First, in this study, we applied collapsed Gibbs samplers to perform posterior inference for the R-IB and conventional models. However, in general, collapsed Gibbs samplers are relatively expensive in terms of computational cost. Thus, one direction is to enhance the scalability and the computational efficiency of the R-IB model, whereas the other is to enhance the capability of the model. Second, the R-IB model can be viewed as an extension of the IRM. Hence, objects within relational data are partitioned into non-overlapping clusters. However, cluster structure underlying real-world relationships could be more complicated. In the future, we intend to extend the R-IB model to be able to discover more advanced structures, such as those with mixed membership [1, 19, 73] or multiple membership [43, 54, 56] assumptions.

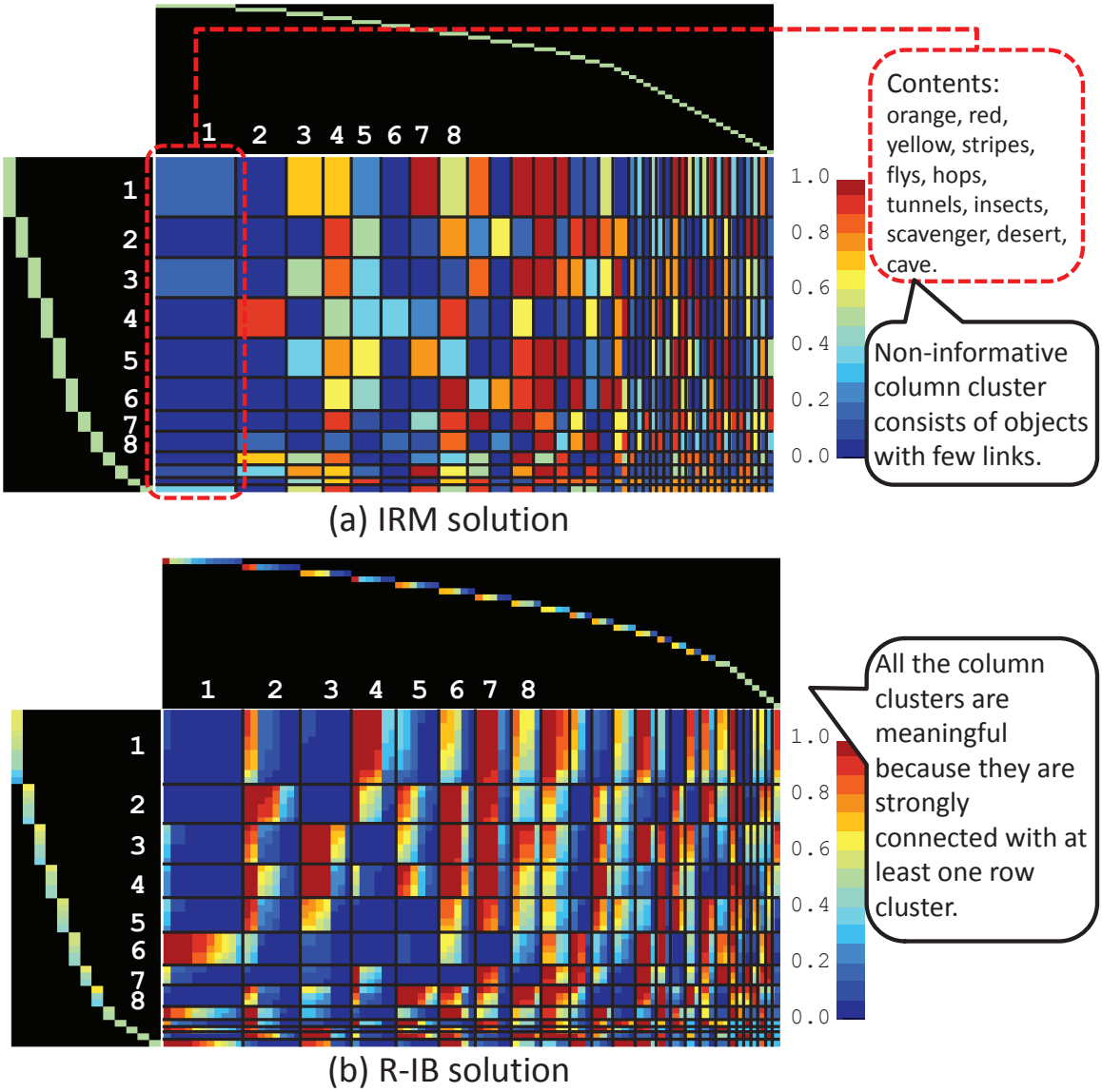


Figure 5.4: Clustering results on the Animal dataset. The central color map denotes estimated link probabilities.

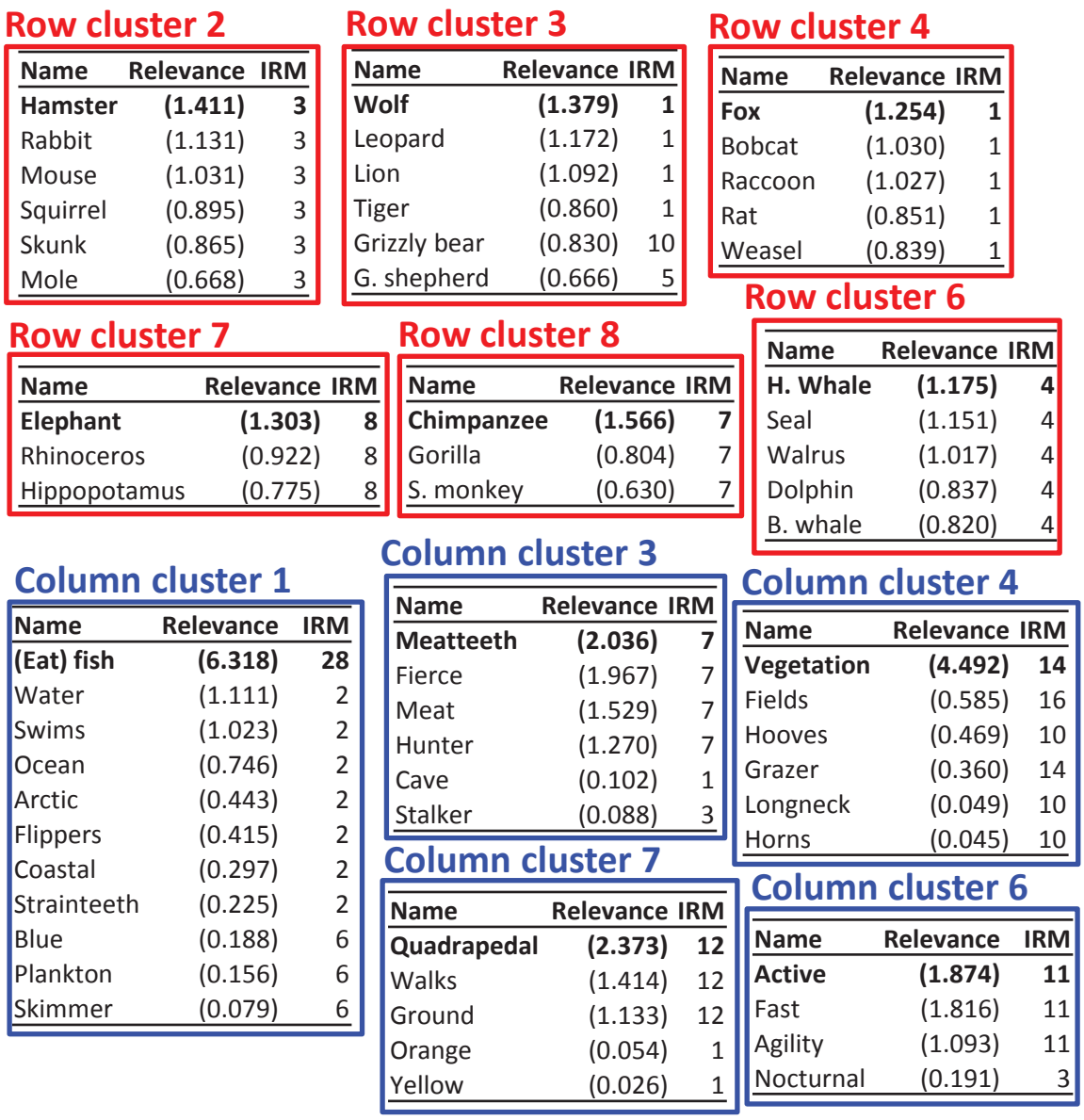


Figure 5.5: Example of clusters obtained by the R-IB with the Animal dataset. Objects within each cluster are sorted in descending order of estimated relevance values. For each object, we list the cluster index that the IRM estimated for the corresponding object (third column). The most relevant object within each cluster is highlighted in bold.

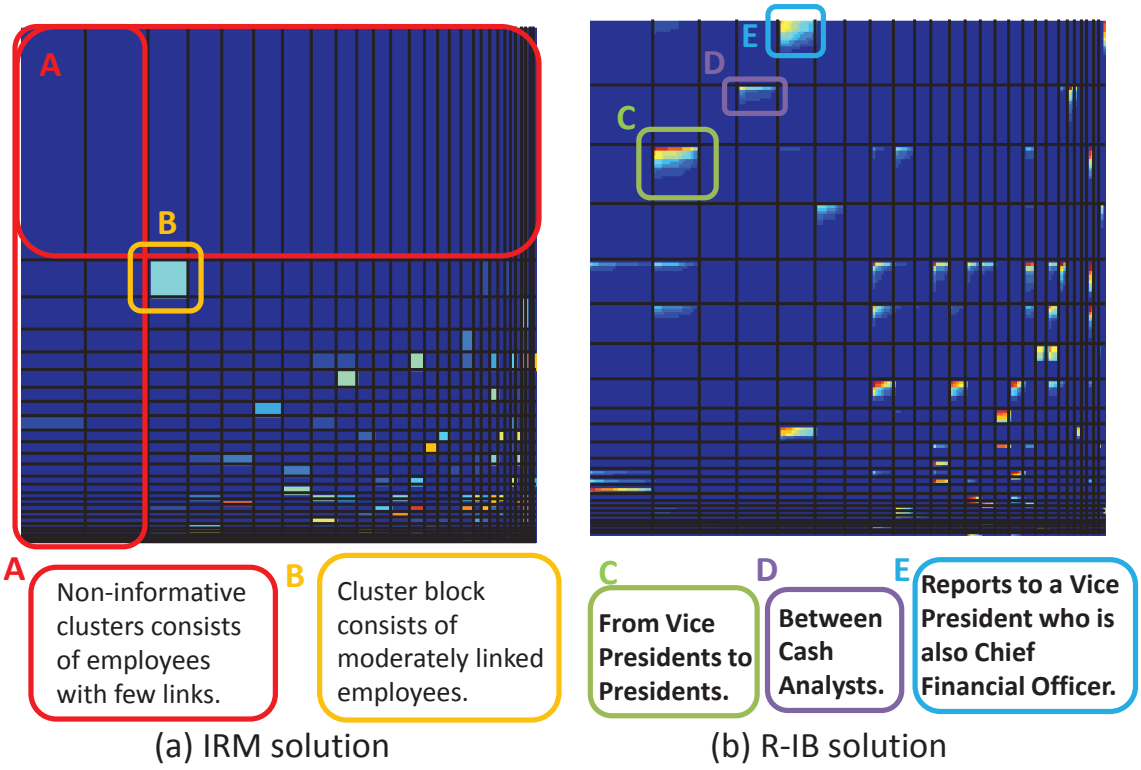


Figure 5.6: Clustering results on the Enron09 dataset.

Chapter 6

Conclusion

In this thesis, we addressed the problem of analyzing relational data while taking account of the relevance of objects. We introduced the relevance-dependent biclustering problem, which simultaneously estimates the bicluster structure and the relevance of objects. Then, we explored several approaches for modeling relevance-dependency and developed new relevance-dependent biclustering models for each approach.

In Chapter 3, we discussed the relevance-dependency modeling using Boolean functions. We introduced a latent binary variable that indicates whether each observation relates to the block structure (foreground distribution) or to a background noise (background distribution). Then, we introduced a mechanism that the binary variable for an entry is determined by calculating an arbitrary Boolean function of Bernoulli trials from row and column objects. By incorporating the mechanism, we proposed a new relevance-dependent biclustering model termed the RDIRM, which can automatically estimate the number of clusters.

In Chapter 4, we generalized the relevance modeling in the RDIRM. By considering continuous relaxation of the Boolean function in the RDIRM, we proposed a mixed-membership mechanism that contains all the Boolean functions as special cases. In the mixed-membership approach, we resolved two critical limitations of relevance modeling

in the RDIRM. First, in the mixed-membership mechanism, the form of the relaxed Boolean function can be automatically estimated from given data. Furthermore, in the mixed-membership mechanism, we can straightforwardly consider relevance values with three or more dimensions. Therefore, we can introduce multiple background distribution for considering different types of irrelevant objects. By incorporating the mixed-membership mechanism, we proposed the MLIRM, which has two background distributions. The relevance parameters in the MLIRM can explain, not only passive objects with few links, but also spamming objects with extremely many links.

In Chapter 5, we introduced a link function approach for modeling relevance-dependency. We introduced the Relevance-dependent Bernoulli Distribution (R-BD), which is a novel prior distribution for relevance-dependent binary matrices. In the R-BD, a link strength for an entry is defined by three non-negative parameters: a typical link strength common to all entries in the matrix, and two relevance parameters for each row and column objects. Then, an observed link probability is directly calculated by transforming the product of these three non-negative variables into a probability using Bernoulli-Poisson link function. The main advantages of the R-BD is as follows. First, the relevance-modeling in the R-BD do not have to consider any background distributions. Thus, the number of latent variables to be estimated is significantly smaller than those in the RDIRM and MLIRM. Second, the link probability in the R-BD can be modulated widely from 0.0 to 1.0 without introducing complicated mechanism as in the MLIRM. Thus, the effect relevance values in the R-BD is interpretable. Finally, as the all parameters of the R-BD can be completely marginalized out, we do not have to explicitly estimate R-BD's parameters when performing posterior inference. By incorporating the R-BD as a component distribution, we proposed a novel biclustering model termed the R-IB. Thanks to the property of the R-BD, the posterior inference for the R-IB can also be performed using a collapsed Gibbs sampler. Furthermore, the R-IB can be inferred faster than not only the RDIRM and MLIRM, but also the

original IRM.

Through this study, we succeeded in opening the beginning of relevance-dependent biclustering research. In this research, we applied our relevance-dependency models to biclustering problem, where objects in relational data are partitioned into non-overlapping clusters. However, mixed-membership or multiple membership assumptions are appropriate in many real-world situations. Therefore, for future work, we have a plan to consider relevance-dependency modeling for more general machine learning problems such as matrix factorization[1].

[1]We have published a result in [54].

Bibliography

[1] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[2] David J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin Heidelberg, 1985.

[3] Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[4] Miguel Araujo, Stephan Günnemann, Gonzalo Mateos, and Christos Faloutsos. Beyond blocks: Hyperbolic community detection. In *Proceedings, Part I, of the 2014 European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2014)*, volume 8724 of *Lecture Notes in Computer Science*, pages 50–65. Springer Berlin Heidelberg, 2014.

[5] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dharmendra S. Modha. A generalized maximum entropy approach to Bregman coclustering and matrix approximation. *Journal of Machine Learning Research*, 8:1919–1986, 2007.

[6] David Blackwell and James B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] John F. Canny. GaP: a factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 122–129. ACM, 2004.

[9] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, pages 93–103. AAAI Press, 2000.

[10] Alonzo Clifford Cohen, Jr. Estimating the parameter in a conditional Poisson distribution. *Biometrics*, 16(2):203–211, 1960.

[11] Yulai Cong, Bo Chen, Hongwei Liu, and Mingyuan Zhou. Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pages 864–873. PMLR, 2017.

[12] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pages 233–240. Omnipress, 2006.

[13] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pages 269–274. ACM, 2001.

[14] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003)*, pages 89–98. ACM, 2003.

[15] Chris H. Q. Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 126–135. ACM, 2006.

[16] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[17] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[18] Rana Forsati, Mehrnoush Mahdavi, Mehrdad an Shamsfard, and Mohamed Sarwat. Matrix factorization with explicit trust and distrust side information for improved social recommendation. *ACM Transactions on Information Systems*, 32(4):17:1–17:38, 2014.

[19] Wenjie Fu, Le Song, and Eric P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, pages 329–336. Omnipress, 2009.

[20] Thomas George and Srujana Merugu. A scalable collaborative filtering framework based on co-clustering. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pages 625–628. IEEE, 2005.

[21] Charles J. Geyer. Lower-truncated Poisson and negative binomial distributions. Technical report, Working Paper Written for the Soft-

ware R. University of Minnesota, MN (available: http://cran.r-project.org/web/packages/aster/vignettes/trunc.pdf), 2007.

[22] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS 2005)*, volume 18 of *Advances in Neural Information Processing Systems*, pages 475–482. MIT Press, 2005.

[23] John A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[24] Qirong Ho, Ankur P. Parikh, Le Song, and Eric P. Xing. Multiscale community blockmodel for network exploration. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, volume 15 of *Proceedings of Machine Learning Research*, pages 333–341. PMLR, 2011.

[25] Peter D. Hoff. Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics*, 61(4):1027–1036, 2005.

[26] Peter D. Hoff. Model-based subspace clustering. *Bayesian Analysis*, 1(2):321–344, 2006.

[27] Changwei Hu, Piyush Rai, and Lawrence Carin. Zero-truncated Poisson tensor factorization for massive binary tensors. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 375–384. AUAI Press, 2015.

[28] Changwei Hu, Piyush Rai, Changyou Chen, Matthew Harding, and Lawrence Carin. Scalable Bayesian non-negative tensor factorization for massive count data. In *Proceedings, Part II, of the 2015 European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD*

2015), volume 9285 of *Lecture Notes in Computer Science*, pages 53–70. Springer International Publishing, 2015.

[29] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[30] Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua B. Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS 2010)*, volume 23 of *Advances in Neural Information Processing Systems*, pages 919–927. Curran Associates, Inc., 2010.

[31] Katsuhiko Ishiguro, Issei Sato, Masahiro Nakano, Akisato Kimura, and Naonori Ueda. Infinite plaid models for infinite bi-clustering. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*, pages 1701–1708. AAAI Press, 2016.

[32] Katsuhiko Ishiguro, Issei Sato, and Naonori Ueda. Averaged collapsed variational Bayes inference. *Journal of Machine Learning Research*, 18(1):1–29, 2017.

[33] Katsuhiko Ishiguro, Naonori Ueda, and Hiroshi Sawada. Subset infinite relational models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, volume 22 of *Proceedings of Machine Learning Research*, pages 547–555. PMLR, 2012.

[34] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence (AAAI 2006)*, volume 1, pages 381–388. AAAI Press, 2006.

[35] Bryan Klimat and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 2004 European Conference on Machine*

Learning (ECML 2004), volume 3201 of *Lecture Notes in Computer Science*, pages 217–226. Springer Berlin Heidelberg, 2004.

[36] Yuval Kluger, Ronen Basri, Joseph T. Chang, and Mark Gerstein. Spectral biclustering of microarray cancer data: Co-clustering genes and conditions. *Genome Research*, 13(4):703–716, 2003.

[37] Takuya Konishi, Takatomi Kubo, Kazuho Watanabe, and Kazushi Ikeda. Variational Bayesian inference algorithms for infinite relational model of network data. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2176–2181, 2015.

[38] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2000.

[39] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems (NIPS 2000)*, volume 13 of *Advances in Neural Information Processing Systems*, pages 556–562. MIT Press, 2000.

[40] Wu-Jun Li and Dit-Yan Yeung. Relation regularized matrix factorization. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1126–1131. IJCAI Organization / AAAI Press, 2009.

[41] Jun S. Liu. The collapsed Gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[42] Jundao Liu, Caihua Wu, and Wenyu Liu. Bayesian probabilistic matrix factorization with social relations and item contents for recommendation. *Decision Support Systems*, 55(3):838–850, 2013.

[43] Morten Mørup, Mikkel N. Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. In *Proceedings of the 2011 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2011)*, pages 1–6. IEEE, 2011.

[44] Masahiro Nakano, Katsuhiko Ishiguro, Akisato Kimura, Takeshi Yamada, and Naonori Ueda. Rectangular tiling process. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, volume 32 of *Proceedings of Machine Learning Research*, pages 361–369. PMLR, 2014.

[45] David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

[46] Michael A. Newton. Discovering combinations of genomic aberrations associated with cancer. *Journal of the American Statistical Association*, 97(460):931–942, 2002.

[47] Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[48] Iku Ohama, Hiromi Iida, Takuya Kida, and Hiroki Arimura. An extension of the infinite relational model incorporating interaction between objects. In *Proceedings, Part II, of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2013)*, volume 7819 of *Lecture Notes in Computer Science*, pages 147–159. Springer Berlin Heidelberg, 2013.

[49] Iku Ohama, Hiromi Iida, Takuya Kida, and Hiroki Arimura. The relevance dependent infinite relational model for discovering co-cluster structure from relation-

ships with structured noise. *IEICE Transactions on Information and Systems*, E99-D(4):1139–1152, 2016.

[50] Iku Ohama, Takuya Kida, and Hiroki Arimura. Multi-layered framework for modeling relationships between biased objects. In *Proceedings of the 15th SIAM International Conference on Data Mining (SDM 2015)*, pages 819–827. SIAM, 2015.

[51] Iku Ohama, Takuya Kida, and Hiroki Arimura. Discovering relevance-dependent bicluster structure from relational data. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 2578–2584. IJCAI Organization, 2017.

[52] Iku Ohama, Takuya Kida, and Hiroki Arimura. Discovering co-cluster structure from relationships between biased objects. *IEICE Transactions on Information and Systems*, E101-D(12):3108–3122, 2018.

[53] Iku Ohama, Takuya Kida, and Hiroki Arimura. Discovering relevance-dependent bicluster structure from relational data: A model and algorithm. *Transactions of the Japanese Society for Artificial Intelligence*, 33(6):B–I46_1–10, 2018.

[54] Iku Ohama, Issei Sato, Takuya Kida, and Hiroki Arimura. On the model shrinkage effect of gamma process edge partition models. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, volume 30 of *Advances in Neural Information Processing Systems*, pages 396–404. Curran Associates, Inc., 2017.

[55] Daniel N. Osherson, Joshua Stern, Ormond Wilkie, Michael Stob, and Edward E. Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991.

[56] Konstantina Palla, David A. Knowles, and Zoubin Ghahramani. An infinite latent attribute model for network data. In *Proceedings of the 29th International Confer-*

ence on Machine Learning (ICML 2012), pages 1607–1614. icml.cc / Omnipress, 2012.

[57] Daniel M. Roy, Charles Kemp, Vikash K. Mansinghka, and Joshua B. Tenenbaum. Learning annotated hierarchies from relational data. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS 2006)*, volume 19 of *Advances in Neural Information Processing Systems*, pages 1185–1192. MIT Press, 2006.

[58] Daniel M. Roy and Yee Whye Teh. The Mondrian process. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008)*, volume 21 of *Advances in Neural Information Processing Systems*, pages 1377–1384. Curran Associates, Inc., 2008.

[59] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS 2007)*, volume 20 of *Advances in Neural Information Processing Systems*, pages 1257–1264. Curran Associates, Inc., 2007.

[60] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, volume 307 of *ACM International Conference Proceeding Series*, pages 880–887. ACM, 2008.

[61] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[62] M. Mahdi Shafiei and Evangelos E. Milios. Latent Dirichlet co-clustering. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006)*, pages 542–551. IEEE, 2006.

[63] Hanhuai Shan and Arindam Banerjee. Bayesian co-clustering. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, pages 530–539. IEEE, 2008.

[64] Ajit Paul Singh and Geoffrey J. Gordon. A Bayesian matrix factorization model for relational data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 556–563. AUAI Press, 2010.

[65] Jagbir Singh. A characterization of positive Poisson distribution and its statistical application. *SIAM Journal on Applied Mathematics*, 34(3):545–548, 1978.

[66] Ilya Sutskever, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Modelling relational data using Bayesian clustered tensor factorization. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009)*, volume 22 of *Advances in Neural Information Processing Systems*, pages 1821–1828. Curran Associates, Inc., 2009.

[67] Koh Takeuchi, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada. Non-negative multiple matrix factorization. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pages 1713–1720. IJCAI Organization / AAAI Press, 2013.

[68] Koh Takeuchi, Ryota Tomioka, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada. Non-negative multiple tensor factorization. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013)*, pages 1199–1204. IEEE, 2013.

[69] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[70] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS 2006)*, volume 19 of *Advances in Neural Information Processing Systems*, pages 1353–1360. MIT Press, 2006.

[71] Pu Wang, Carlotta Domeniconi, and Kathryn B. Laskey. Latent Dirichlet Bayesian co-clustering. In *Proceedings, Part II, of the 2009 European Conference on Machine Learning & Principles and Knowledge Discovery in Databases (ECML-PKDD 2009)*, volume 5782 of *Lecture Notes in Computer Science*, pages 522–537. Springer Berlin Heidelberg, 2009.

[72] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Infinite hidden relational models. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 544–551. AUAI Press, 2006.

[73] Mingyuan Zhou. Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2014)*, volume 27 of *Advances in Neural Information Processing Systems*, pages 3455–3463. Curran Associates, Inc., 2014.

[74] Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*, volume 38 of *Proceedings of Machine Learning Research*, pages 1135–1143. PMLR, 2015.

[75] Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.

[76] Mingyuan Zhou, Yulai Cong, and Bo Chen. The Poisson gamma belief network. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS 2015)*, volume 28 of *Advances in Neural Information Processing Systems*, pages 3043–3051. Curran Associates, Inc., 2015.

[77] Mingyuan Zhou, Lauren Hannah, David B. Dunson, and Lawrence Carin. Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471. PMLR, 2012.