Discovering Co-Cluster Structure from Relationships between Biased Objects

Iku OHAMA†, Nonmember, Takuya KIDA††, and Hiroki ARIMURA††, Members

SUMMARY  Latent variable models for relational data enable us to extract the co-cluster structure underlying observed relational data. The Infinite Relational Model (IRM) is a well-known relational model for discovering co-cluster structures with an unknown number of clusters. The IRM and several related models commonly assume that the link probability between two objects depends only on their cluster assignment. However, relational models based on this assumption often lead us to extract many non-informative and unexpected clusters. This is because the cluster structures underlying real-world relationships are often blurred by biases of individual objects. To overcome this problem, we propose a multi-layered framework, which extracts a clear de-blurred co-cluster structure in the presence of object biases. Then, we propose the Multi-Layered Infinite Relational Model (MLIRM) which is a special instance of the proposed framework incorporating the IRM as a co-clustering model. Furthermore, we reveal that some relational models can be regarded as special cases of the MLIRM. We derive an efficient collapsed Gibbs sampler to perform posterior inference for the MLIRM. Experiments conducted using real-world datasets have confirmed that the proposed model successfully extracts clear and interpretable cluster structures from real-world relational data.

1. Introduction

Relational data encoding pairwise relationships between $N$ objects and $M$ objects appear in many fields. For example, the point-of-sale (POS) data of an e-commerce (EC) site is relational data between customers and items; follower lists in social networking services (SNS), such as Twitter, is relational data among users. In this paper, we focus on the most typical type of relational data; i.e., sets of observations represented as a binary matrix. One of the major motivations behind analyzing such relational data is to discover the hidden structures of interaction between objects. For example, someone might want to obtain the following knowledge:

- How should we segment the customers of an EC site in order to understand their preferences more clearly?
- How many roles are there among the employees of a company and which employees have the same role?

Many statistical models that can answer such inquiries have been proposed [1]–[5]. In particular, the Stochastic Block Model (SBM) [4] and the Infinite Relational Model (IRM) [5] are well-known co-clustering models for relational data. Using these models, relationships between objects can be abstracted to relationships between clusters. The SBM requires specifying the number of clusters, whereas the IRM automatically estimates the number of clusters depending on the observed data. Because the SBM and IRM are so popular, many extensions have been proposed for these models [6]–[14]. In SBM, IRM, and their extended models, each object in relational data has a latent variable indicating its cluster membership. Then, these models assume that a link probability between two individual objects depends only on their cluster memberships (Fig. 1 (a)).

However, when analyzing real-world relational data, these IRM families frequently discover many non-informative and unexpected clusters. This is because, in real-world relationships, the underlying cluster structure is often blurred. For example, when clustering the relationships between customers and items from purchase records, one of the primary requirements is that the obtained cluster structures indicate the preference types of customers or the functional categories of items. However, the link proba-
bility between a customer and an item depends not only on the customer’s preference and the item’s function but also on other reasons. For example, when compared to another customer with a larger budget, the link probability related to a customer with a smaller budget decreases regardless of their preference. Yet another example is a very popular item, such as a best-selling book, which may be purchased by many customers. In this case, the link probability related to the item increases regardless of its content; i.e., an object in real-world relational data has its own bias, which blurs the underlying cluster structure (Fig. 1 (b)). To obtain an essential co-cluster structure in such situations, it is natural to assume additional latent factors that affect the link probabilities of objects regardless of their cluster membership.

To overcome these problems, 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. In this paper, 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.

The main contributions of this paper are as follows:

- We introduce a new multi-layered framework that simultaneously extracts clear co-cluster structure and objects’ bias from blurred relational data (Sect. 3.1).
- By incorporating the IRM to our framework, we propose a new co-clustering model called the MLIRM (Sect. 3.2).
- We reveal the relationships between the proposed MLIRM and various conventional relational models (Sect. 3.3).
- To perform posterior inference for the MLIRM, we derive an efficient collapsed Gibbs sampler (Sect. 4).
- We conduct experiments using several real-world datasets (Sect. 5). In quantitative comparison, we confirm that the MLIRM finds simpler co-cluster structure with better generalization ability than conventional models. Furthermore, we qualitatively show that the MLIRM solutions allow deep insights to be gained from real-world relational data.

The remainder of this paper is organized as follows. In Sect. 2, we review the SBM and its families as baselines for co-clustering models and discuss their drawbacks. In Sect. 3, we describe the generative processes of the proposed framework and the MLIRM. We also explain the relationship between the MLIRM and several recently proposed models. In Sect. 4, we derive the Gibbs sampling algorithm for performing posterior inference for the MLIRM. Section 5 discusses experimental evaluations using synthetic and real-world datasets. In Sect. 6, we offer suggestions for future work and conclude the paper.

A portion of this work was presented at The 15th SIAM International Conference on Data Mining (SDM 2015) [15]. This paper is a full version of [15]. In response to comments we received from SDM 2015 reviewers, we refined the description and added several diagrams in order to clarify our motivation. Moreover, we provide detailed derivations that were omitted in the conference paper due to space limitations. Furthermore, we added figures for our MLIRM solution in order to allow the contents of the extracted cluster to be assessed.

2. Stochastic Block Models

2.1 Relational Data

Here, we define the relational data discussed in this paper. Let \( R \) be the \( N^1 \times N^2 \) binary matrix that represents relational data between sets of objects \( T^1 = \{O_{11}^1, \ldots, O_{1N^1}^1\} \) and \( T^2 = \{O_{11}^2, \ldots, O_{1N^2}^2\} \). An entry \( R_{i,j} = 1(0) \) indicates that there is a link (non-link) between \( O_{i}^1 \) and \( O_{j}^2 \). 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).

Several variations of relational data can be considered straightforwardly. For example, discrete-valued relational data are preferable for encoding customer ratings for items on an EC site. As another example, continuous-valued relational data can be used to represent relationships on sensor networks. In addition, we can consider relational data encoding relationships among three or more domains (e.g., customers \( \times \) items \( \times \) time) by tensor representation. Although these variants are important, we focus on two-domain binary relational data, which is the most typical type of relational data.
Latent variable models for co-clustering consider a stochastic distribution over $R$. The SBM [4], the IRM [5], and the Mixed Membership Stochastic Block Model (MMSBM) [6] are popular among such models. Let $K$ and $L$ be the number of clusters for $T^1$ and $T^2$, respectively. In general, these models have latent variables $c^1_i = \{c^1_i|_{k=1}^K\} \in [0, 1]^K$ and $c^2_j = \{c^2_j|_{l=1}^L\} \in [0, 1]^L$ for $T^1$ and $T^2$, respectively. Note that $\sum_{i=1}^K c^1_i = \sum_{l=1}^L c^2_j = 1$. These latent variables indicate the cluster membership for each object. In addition, let $\eta_k$ be the $K \times L$ matrix indicating 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 a row object assigned to cluster $k$ and a column object assigned to cluster $l$. Thus, the link probability between $O^1_i$ and $O^2_j$ is given as follows:

$$P(R_{i,j} = 1 | c^1_i, c^2_j, \eta) = c^1_i^{\top} \eta c^2_j = \sum_k \sum_l c^1_i c^2_j \eta_{k,l}. \quad (1)$$

The above model is exactly equivalent to the MMSBM. If the latent variables $c^1_i$ and $c^2_j$ are constrained to have hard assignments as $c^1_i = \{c^1_i|_{k=1}^K\} \in \{0, 1\}^K$ and $c^2_j = \{c^2_j|_{l=1}^L\} \in \{0, 1\}^L$, the above model is equivalent to the SBM and IRM. Figure 3 shows the diagram of SBMs’ solution. As can be seen, SBMs factorize given relational data into three low dimensional matrices. The only difference between the SBM and IRM is the prior distribution for cluster assignments. In the IRM, the Chinese Restaurant Process (CRP) [16], [17] is used as a prior for cluster assignments of an unknown number of clusters, while a fixed dimensional Dirichlet distribution is used for the SBM. By estimating latent variables, a cluster structure can be obtained from observed relational data.

There are several approaches to perform posterior inference for the SBM family. In particular, the Gibbs sampler and the variational Bayes inference are frequently used. The Gibbs sampler guarantees asymptotic convergence to the true posterior, whereas the variational Bayes inference is an approximative approach. In this study, we apply an improved Gibbs sampler called the collapsed Gibbs sampler [18] for inferring the posterior. In the collapsed Gibbs sampler, some of the model parameters are marginalized out. Therefore, we need to sequentially update only the remaining latent variables, which leads to faster convergence compared to the naive Gibbs sampler.

### 2.2 Stochastic Block Models

Due to underlying model assumptions, these SBM families often discover many unexpected clusters when analyzing real-world relational data. As evidenced by the form of Eq. (1), these models commonly assume that link probability between objects depends only on their cluster assignments ($c^1_i$ and $c^2_j$). However, this assumption is not acceptable for some real-world situations because each object often has its own bias. For example, when analyzing the relationships between customers and items in purchase records, one particular interest is to extract the types of customer preferences. However, in this case, the probability of a customer purchasing an item depends not only on the customer’s preference for the item but also on their biases, such as budget constraints, frequency of visiting the shopping site, and visibility of the item. Therefore, the cluster structure underlying the real-world relational data is often blurred by such biases. Consequently, applying the SBM (or one of its variants) to blurred data results in the extraction of many non-informative clusters.

### 3. Multi-Layered Framework

Here, we propose a novel framework for modeling blurred relational data. We also 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.

#### 3.1 Framework Description

To capture the cluster structure and object biases independently, we propose a multi-layered framework. In the proposed framework, we assume that each observation within relational data comes from a mixture distribution of three layers: a clustering layer $c^1_i^{\top} \eta c^2_j$, a background link layer $\eta^1$ and a background non-link layer $\eta^0$. 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 \( \eta^1 \) and \( \eta^0 \).

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 \( \theta_i^1 = (\theta_i^1, \theta_i^0, \theta_i^1) \) and \( \theta_j^2 = (\theta_j^2, \theta_j^0, \theta_j^2) \) for each object, where \( fg, bg_1, \) and \( bg_0 \) indicate the clustering layer, the background link layer, 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 \( w_{i,j} = (w_{i,j|fg}, w_{i,j|bg_1}, w_{i,j|bg_0}) \) be the mixture ratio of each layer for an entry \( R_{i,j} \). Then, we consider the interaction mechanism that transforms two bias parameters \( \theta_i^1 \) and \( \theta_j^2 \) into a mixture ratio \( w_{i,j} \). 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 \( \theta_i^1, \theta_j^2 \) and the corresponding mixture ratio \( w_{i,j} \). For example, when an bias parameter for the \( i \)-th row object leans to the foreground layer \( (\theta_i^1 \rightarrow 1.0) \), the mixture ratio related to the \( i \)-th row object should also lean to the foreground \( (w_{i,j|fg} \rightarrow 1.0) \). However, the form of transformation \( f : \theta_i^1, \theta_j^2 \rightarrow w_{i,j} \) is non-trivial. Thus, we introduce a constrained \( 3 \times 3 \times 3 \) interaction weight \( \phi \) in order to describe the general form of the interaction between two biases. Using indices \( s, t, u \in \{fg, bg_1, bg_0\} \), the mechanism is given as follows:

\[
\begin{align*}
    w_{i,j|u} &= \theta_i^1 \phi_{s,t,u} \theta_j^2 = \sum_{s' \in \{fg, bg_1, bg_0\}} \theta_i^{s'}, \theta_j^{s'}, \phi_{s,t,u} \\
    \text{s.t.} & \quad \sum_{s'} \phi_{s,t,u} = 1, \quad \phi_{s,t,u} \in [0, 1] \\
    & \quad \phi_{s,t,u} = 0.0 \text{ if } (u \neq s \text{ and } u \neq t) \forall s, t, u, \quad (2)
\end{align*}
\]

where \( \phi_{s,t,u} \) is a matrix that is a slice of \( \phi \) related to \( u \). The constraints in Eq. (2) ensure a positive correlation between \( \theta_i^1, \theta_j^2 \) and \( w_{i,j} \). More intuitively, the constraints can be understood as follows:

- If row object \( i \) and column object \( j \) select the same layer according to \( \theta_i^1 \) and \( \theta_j^2 \), 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 \( \phi \). Finally, \( R_{i,j} \) is generated from the selected layer.

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

\[
P(R_{i,j} = 1 | \theta_i^1, \theta_j^2, \phi) = \left( \begin{array}{ccc}
    \sum_{u} w_{i,j|u} c_i^T \eta c_j^2 & \eta^1 \eta^0 \\
    \sum_{u} w_{i,j|u} c_j^T \eta c_i^2 & \eta^1 \eta^0 \\
    \eta^1 \eta^0 & \eta^1 \eta^0
\end{array} \right)
\]

where \( \theta_i^1 \phi_j^2 = \{ \theta_i^1 \phi_j^2 \}_{j \in \{fg, bg_1, bg_0\}} \).

To give the hierarchical generative model for the proposed framework, we use the extended definition of the Dirichlet distribution presented by Ferguson [19]. 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 \( \phi \) with the positive correlation constraint given by Eq. (2) can be constructed using Dirichlet distributions with constrained parameters. In summary, the hierarchical generative model of the multi-layered framework is as follows:

\[
\begin{align*}
    \theta_i^1 | \alpha^1 & \sim \text{Dir}(\alpha^1), \quad \theta_j^2 | \alpha^2 \sim \text{Dir}(\alpha^2), \\
    \phi_{s,t,u} | a_{s,t} & \sim \text{Dir}(a_{s,t}) \quad \text{s.t. } a_{s,t,u} = 0.0 \text{ if } u \neq s \text{ and } u \neq t, \\
    R_{i,j} | c_i^T \eta c_j^2, \eta^1, \eta^0, \theta_i^1, \theta_j^2, \phi & \sim \text{Bern} \left( \theta_i^1 \phi_j^2 \right) \cdot \left( \begin{array}{c}
    c_i^T \eta c_j^2 \\
    \eta^1 \\
    \eta^0
\end{array} \right), \quad (6)
\end{align*}
\]

where \( \text{Dir}(\cdot) \) is the Dirichlet distribution, and \( \text{Bern}(\cdot) \) is the Bernoulli distribution. Figure 4 presents a diagram of the proposed multi-layered framework. Now, we briefly review the above mentioned process. First, the bias parameters \( \theta_i^1 \) and \( \theta_j^2 \) for each object \( O_i \) and \( O_j \) are given from Dirichlet priors with parameters \( \alpha^1 = \{ \alpha^1_{i,j} \}_{i,j \in \{fg, bg_1, bg_0\}} \) and \( \alpha^2 = \{ \alpha^2_{i,j} \}_{i,j \in \{fg, bg_1, bg_0\}} \), respectively (Eq. (4)). Second, an interaction weight \( \phi = \{ \phi_{s,t,u} \}_{s,t,u \in \{fg, bg_1, bg_0\}} \) is given by Dirichlet distributions with constrained hyperparameters \( a_{s,t} \) (Eq. (5)). Finally, the link \( R_{i,j} \) is generated by a Bernoulli distribution (Eq. (6)).

3.2 The Multi-Layered Inferential Relational Model

Here, we propose a new generative model called the MLIRM, which is a concrete instance of the multi-layered framework proposed in Sect. 3.1. In the MLIRM, the IRM is embedded as a prior for the clustering model \( c_i^T \eta c_j^2 \) given in Eq. (6). The IRM is a hard clustering model, i.e., an object is assigned to a single cluster. Hereafter, let \( z_i^1 \in \{1, \cdots, K\} \) and \( z_j^2 \in \{1, \cdots, L\} \) be the cluster indices to which \( O_i \) and \( O_j \) are assigned, respectively. Therefore, in the IRM, the link probability of \( R_{i,j} \) is given by \( \eta_{k,l} \), where \( z_i^1 = k \) and \( z_j^2 = l \). To estimate the number of clusters automatically from the given data, the IRM uses the CRP as the prior distribution for \( z_i^1 \) and \( z_j^2 \). To summarize, the full description of the generative model for the MLIRM is as follows:

\[
\begin{align*}
    z_i^1 | \gamma^1 & \sim \text{CRP}(\gamma^1), \quad z_j^2 | \gamma^2 \sim \text{CRP}(\gamma^2), \quad (7)
\end{align*}
\]
Fig. 4 (Best viewed in color.) Multi-layered framework diagram.

$$
\begin{align*}
\eta_{k,l} | \beta & \sim \text{Beta}(\beta, \beta), \\
\eta^1 | \beta & \sim \text{Beta}(\beta, \beta), \quad \eta^0 = 1 - \eta^1, \\
\theta^1_i | \alpha^1 & \sim \text{Dir}(\alpha^1), \quad \theta^2_j | \alpha^2 \sim \text{Dir}(\alpha^2), \\
\phi_{s,t} | a_{s,t} & \sim \text{Dir}(a_{s,t}), \\
\text{s.t. } a_{s,t} & = 0.0 \text{ if } u \neq s \text{ and } u \neq t,
\end{align*}
$$

where Beta(·,·) is the beta distribution. Figure 5(a) shows a graphical representation for the MLIRM. Equations (7) and (8) are the embedded IRM for the clustering layer of the MLIRM. Note that $\gamma^1$ and $\gamma^2$ are the concentration parameters for the CRPs. Equation (9) defines the background layers. In the MLIRM, we define $\eta^1 = 1 - \eta^0$ in order to ensure that $\eta^1$ and $\eta^0$ capture irrelevant links and irrelevant non-links, respectively.
3.3 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. [5] 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 $O_i \cdot O_j$ depends only on their corresponding cluster assignments $z_i$ and $z_j$. Then, the link $R_{i,j}$ is drawn as follows:

$$R_{i,j} \mid z_i, z_j, \eta \sim \text{Bern}(\eta_{z_i, z_j}).$$

(13)

Note that the IRM is a special case of the proposed MLIRM when the bias parameters $\theta_1 = (\theta_{1\gamma, \eta}, \theta_{1\gamma, \eta}, \theta_{1\gamma, \eta})^T$ and $\theta_2 = (\theta_{2\gamma, \eta}, \theta_{2\gamma, \eta}, \theta_{2\gamma, \eta})^T$ 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 [20], [21] relative to clustering biological sequences. For co-clustering relational data, there have been several extensions of the IRM so that the background layer affects link probabilities. The Subset IRM (SIRM) proposed by Ishiguro et al. [22] and the Relevance Dependent IRM (RDIRM) proposed by Ohama et al. [23], [24] both consider a generative model in which link probability is a mixture distribution of a clustering layer $\eta_k$ and a background layer $\eta^b$. In the SIRM, binary variables $s_i^1, s_i^2 \in \{0, 1\}$ are introduced to indicate whether object $i$ is relevant to the underlying cluster structure. Then, a subset of $R$, where $s_i^1 \times s_i^2 = 1$, is explained by the clustering model $\eta_k$, while the rest are explained by the background probability $\eta^b$. Specifically, in the SIRM, the link $R_{i,j}$ is drawn as follows:

$$R_{i,j} \mid z_i, z_j, s_i^1, s_i^2, \eta, \eta^b \sim \text{Bern}\left( I(\eta_{z_i, z_j}) \right).$$

(14)

where $I(\cdot)$ 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, the background layer $\eta^b$ is divided into two layers that are related to spamming and passive objects. Specifically, in the RDIRM, the link $R_{i,j}$ is drawn as follows:

$$R_{i,j} \mid z_i, z_j, s_i^1, s_i^2, \eta, \eta^b \sim \text{Bern}\left( \frac{\eta_{z_i, z_j}}{\eta_{z_i, z_j} + \eta_{z_i, z_j}} \right).$$

(15)

The SIRM and RDIRM can be viewed as special cases of the MLIRM. The MLIRM is equivalent to the RDIRM in Eq. (15), if the bias parameters $\theta_1$ and $\theta_2$ are constrained to $\theta_{1\gamma, \eta} = \theta_{2\gamma, \eta} = 0$ and if the interaction tensor $\phi$ is given stochastically. In addition to these constraints, the MLIRM is equivalent to the SIRM in Eq. (14) if $\theta_1, \theta_2 \in \{1, 0, 0\}$ and $\phi_{\eta_{z_i, z_j}} = (0, 0, 1)$.

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 $\eta^b$. Therefore, the estimated value of $\eta^b$ tends to lean either on 0 or 1. Consequently, the RDIRM can only capture either passive or spamming objects. Furthermore, the RDIRM requires specifying the static function $f(\eta, \cdot)$ 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 $\phi$ can be estimated automatically from the data. To summarize, the advantages of the MLIRM are as follows.

- The MLIRM introduces two background layers $\eta^l$ and $\eta^b$ in order to accommodate relational data with both spamming and passive objects.
- The generative process for transformation $f : \theta_1, \theta_2 \rightarrow w_{i,j}$ via $\phi$ enables the interaction structure of two bias parameters to be estimated automatically.

4. Inference

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

4.1 Marginal Likelihood

Thanks to the conjugacy between MLIRM parameters and its prior distributions, $\eta, \eta^l, \eta^b, \theta_1, \theta_2$, and $\phi$ are fully marginalized out.

Introducing three auxiliary variables $r_{i,j}^{\gamma} \sim \text{CRP}(\gamma^{\gamma})$, $r_{i,j}^{\gamma} \sim \text{CRP}(\gamma^{\gamma})$, $\eta_{i,j} \sim \text{Beta}(\beta, \beta)$, $\eta^l \sim \text{Beta}(\beta, \beta)$, $\eta^b = 1 - \eta^l$, $\theta_1 \sim \text{Dir}(\alpha^1)$, $\theta_2 \sim \text{Dir}(\alpha^2)$, $\phi_{\eta_{i,j}} \sim \text{Dir}(\alpha_{\eta_{i,j}})$.

(16)

(17)

(18)

(19)

(20)

s.t. $\alpha_{s,t} = 0.0$ if $u \neq s$ and $v \neq t$.

(21)
where \( \text{Cat}(\cdot) \) is the categorical distribution. The graphical representation of the augmented representation for the MLIRM is depicted in Fig. 5(b). Consequently, we obtain a marginal representation of the augmented representation for the MLIRM as depicted in Fig. 5(c). Following the marginal representation, Eqs. (19)–(23) can be equivalently rewritten as follows:

\[
\begin{align*}
\alpha & \sim \text{DCM}(\alpha^1, \alpha^2), \\
\beta & \sim \text{DCM}(\beta^1, \beta^2), \\
\theta & \sim \text{DCM}(\theta^1, \theta^2), \\
\phi & \sim \text{Bernoulli parameters}. \\
N & = \sum_{i \in T^1} \sum_{j \in T^2} R_{i,j}, \\
M & = \sum_{i \in T^1} \sum_{j \in T^2} (1 - R_{i,j}).
\end{align*}
\]

Second, the terms related to the DCM distribution in Eq. (27) are derived as follows:

\[
P(r \mid r^1, r^2) = \prod_{s} \prod_{t} \frac{\Gamma(\sum_{i} a_{s,ru})}{\Gamma(\sum_{i} a_{s,ru} + n_{s,ru})} \prod_{u} \frac{\Gamma(\alpha^1_{u})}{\Gamma(\alpha^1_{u} + n^1_{u})} \prod_{r} \frac{\Gamma(\alpha^2_{r})}{\Gamma(\alpha^2_{r} + n^2_{r})},
\]

where \( \Gamma(\cdot) \) is the gamma function. The symbols \( n \), \( n^1 \), and \( n^2 \) denote the counts defined for the auxiliary variables \( r \), \( r^1 \), and \( r^2 \), respectively, and are computed as follows:

\[
n_{s,ru} = \sum_{i \in T^1} \sum_{j \in T^2} \mathbb{I}(r_{i,j} = s) \mathbb{I}(r_{j,i} = t) \mathbb{I}(r_{i,j} = u),
\]

where \( \mathbb{I}(\cdot) \) is the indicator function. Finally, the cluster assignments \( z^1 \) and \( z^2 \) follow the CRP. Therefore, we obtain

\[
P(z^1) = \binom{N^1}{n^1} \Gamma(n^1) \prod_{i} \Gamma(m^1_{i}),
\]

\[
P(z^2) = \binom{N^2}{n^2} \Gamma(n^2) \prod_{i} \Gamma(m^2_{i}),
\]

where \( m^1_{i} = \sum_{j} \mathbb{I}(z^1_{j} = k) \) and \( m^2_{i} = \sum_{j} \mathbb{I}(z^2_{j} = l) \).

\subsection{4.2 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 [18], [25]. In this paper, we use the collapsed Gibbs sampler [18] to infer the MLIRM because the algorithm guarantees asymptotic convergence to the true posterior by
4.2.1 Sampling Cluster Assignments \(z^1\) and \(z^2\)

\(z^1\) and \(z^2\) can be sampled in the same way; thus, here, we concentrate only on \(z^1\). Given \(r\), the cluster assignments depend only on the subset of observations, where \(r_{ij} = fg\). Therefore, the conditional posterior for \(z^1\) is derived from Eq. (27) as follows:

\[
P(z^1_{i} = k^* \mid z^1_{-i}, z^2, r, R) \propto m^1_{i,k^*} \times \prod_i \frac{Bl_{m_{i,k^*}+x_{i,k^*}}}{Bl_{m_{i,k^*}+x_{i,k^*}+y_{k^*}}} m^1_{i,k^*} > 0, \gamma^1 \times \prod_i \frac{Bl_{m_{i,k^*}+x_{i,k^*}}}{Bl_{m_{i,k^*}+x_{i,k^*}+y_{k} \neq k^*}} m^1_{i,k^*} = 0, \tag{42}
\]

where \(z^1_{i}\) 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 \(m_{ik}\) and \(\overline{m}_{ik}\) with superscripts are computed as follows:

\[
m^1_{k^*,i} = \sum_{x \in \mathcal{T}} \sum_{j \in \mathcal{T}_1} R_{j,ix} \delta(r_{j,i} = fg) \delta(z^1_{-i} = k^*) \delta(z^2 = l), \tag{43}
\]

\[
\overline{m}^1_{k^*,i} = \sum_{x \in \mathcal{T}} \sum_{j \in \mathcal{T}_1} (1-R_{j,ix}) \delta(r_{j,i} = fg) \delta(z^1_{-i} = k^*) \delta(z^2 = l), \tag{44}
\]

\[
m^1_{k^*,i} = m^1_{k^*,i} + \sum_{j \in \mathcal{T}_1} R_{j,ix} \delta(z^2_{j} = l), \tag{45}
\]

\[
\overline{m}^1_{k^*,i} = \overline{m}^1_{k^*,i} + \sum_{j \in \mathcal{T}_1} (1-R_{j,ix}) \delta(z^2_{j} = l). \tag{46}
\]

4.2.2 Sampling Auxiliary Variables \(r^1\), \(r^2\), and \(r\)

The naive sampler of \(r^1_{-i,j}, r^2_{-i,j}\), and \(r_{ij}\) 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 \(r^1_{-i,j}\) be the grouped variables \(\{r^1_{-i,j}, r^2_{-i,j}, r_{ij}\}\). Then, the conditional posterior is given as follows:

\[
P(r^1_{-i,j} = s^*, r^2_{-i,j} = t^* \mid R, z^1, z^2, r_{-(i,j)}, r^1_{-(i,j)}, r^2_{-(i,j)}) \propto P(r^1_{ij} = u^* \mid r_{-(i,j)}, z^1, z^2, r_{-(i,j)}, r^1_{-(i,j)}, r^2_{-(i,j)}) \times P(r^1_{-i,j} = s^* \mid r^1_{-(i,j)}) \times P(r^2_{-i,j} = t^* \mid r^2_{-(i,j)}). \tag{47}
\]

where \(r^1_{-(i,j)}\) and \(r^2_{-(i,j)}\) denote the entire set of \(r^1\) and \(r^2\) excluding \(r^1_{ij}\) and \(r^2_{ij}\), respectively. Similarly, \(r_{-(i,j)}\) denotes the entire set of \(r\) without \(r_{ij}\). The terms on the right hand side of Eq. (47) are computed as follows:

\[
P(r_{ij} = u^* \mid r_{-(i,j)}, z^1, z^2, r_{-(i,j)}, r^1_{-(i,j)}, r^2_{-(i,j)}) = \begin{cases} \frac{(m_{ij} + \alpha) (m_{ij} + \alpha + 1)}{m_{ij} + m_{ij} + n_{ij} + 2} & u^* = fg, \\ \frac{m_{ij} + \alpha}{m_{ij} + m_{ij} + n_{ij} + 2} & u^* = bg1, \\ \frac{m_{ij} + \alpha}{m_{ij} + m_{ij} + n_{ig} + 2} & u^* = bg0, \end{cases} \tag{48}
\]

\[
P(r^1_{ij} = s^* \mid r_{-(i,j)}, z^1, z^2, r_{-(i,j)}, r^1_{-(i,j)}, r^2_{-(i,j)}) = \frac{n_{s^{*},r^{1,*}}}{\sum_i (n_{s^{*},r^{1,*}} + n_{s^{*},r^{1,*}})}, \tag{49}
\]

\[
P(r^2_{ij} = t^* \mid r^1_{-(i,j)}) \propto \frac{\alpha_{r^{2,*}} + \sum_{j \in \mathcal{T}_1} \delta(r^1_{-i,j} = s^*)}{\alpha_{r^{2,*}} + \sum_{j \in \mathcal{T}_1} \delta(r^1_{-i,j} = s^*)}, \tag{50}
\]

where the related counts are computed as follows:

\[
m_{k^*,i} = m_{k^*,i} - R_{j,ix} \delta(r_{j,i} = fg) \delta(z^1_{-i} = k^*) \delta(z^2 = l), \tag{51}
\]

\[
m_{k^*,i} = m_{k^*,i} - R_{j,ix} \delta(r_{j,i} = bg1), \tag{52}
\]

\[
m_{k^*,i} = m_{k^*,i} - R_{j,ix} \delta(r_{j,i} = bg0), \tag{53}
\]

\[
m_{k^*,i} = m_{k^*,i} - (1 - R_{j,ix}) \delta(z^1_{-i} = k^*) \delta(z^2 = l), \tag{54}
\]

\[
m_{k^*,i} = m_{k^*,i} - m_{bg}, \tag{55}
\]

\[
m_{k^*,i} = m_{k^*,i} - (1 - R_{j,ix}) \delta(r_{j,i} = bg1), \tag{56}
\]

\[
m_{k^*,i} = m_{k^*,i} - m_{bg}, \tag{57}
\]

\[
n_{s^{*},r^{1,*}} = n_{s^{*},r^{1,*}} + \delta(r^1_{-i,j} = s^*) \delta(r^1_{-i,j} = t^*) \delta(r_{ij} = u^*). \tag{58}
\]

4.3 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 [26]–[29] techniques.

4.3.1 Data Augmentation

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

\[
P(N \mid U) \propto \frac{\Gamma(U)}{\Gamma(U + N)}, \tag{59}
\]

\[
P(N' \mid U') \propto \frac{\Gamma(U' + N')}{\Gamma(U')}, \tag{60}
\]

where \(N, N'\) are positive integer random variables and \(U, U'\) are positive real random variables, respectively. Unfortunately, in these cases, we cannot derive straightforward posterior Gibbs samplers for \(U\) and \(U'\) because conjugate priors for Eqs. (59) and (60) have not been developed so far.

To overcome the above-mentioned difficulties, data
augmentation techniques consider expanded joint probabilities over target and auxiliary variables. Let us denote by $\text{Gam}(e_0, f_0)$ the gamma prior with shape parameter $e_0$ and rate parameter $f_0$; i.e., $P(X | e_0, f_0) = \lambda^{e_0-1}e^{-f_0\lambda}/\Gamma(e_0)/f_0^{e_0}$. The key strategies are to use the following expansions:

$$\frac{\Gamma(U)}{\Gamma(U + N)} = \frac{1}{\Gamma(N)} \int_0^1 \! p^{U-1}(1 - p)^{N-1} dp,$$  \hspace{1cm} (61)

$$\frac{\Gamma(U' + N')}{\Gamma(U')} = \sum_{q=1}^N S(N', q) U'^q,$$  \hspace{1cm} (62)

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

By expanding Eq. (59) using Eq. (61), the joint distribution over $N$ and $p$ given $U$ is described as

$$P(N, p | U) \propto \frac{1}{\Gamma(N)} p^{U-1}(1 - p)^{N-1}.$$  \hspace{1cm} (63)

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

$$p | N, U \sim \text{Beta}(U, N),$$  \hspace{1cm} (64)

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 $U \sim \text{Gam}(e_0, f_0)$, the posterior for $U$ is given by

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

Consequently, posterior sampling for $U$ can be performed as

$$U | N, p \sim \text{Gam}(e_0, f_0 - \ln p).$$  \hspace{1cm} (66)

Similarly, by expanding Eq. (60) using Eq. (62), the joint distribution over $N'$ and $q$ given $U'$, is described as

$$P(N', q | U') \propto S(N', q) U'^q,$$  \hspace{1cm} (67)

where the posterior for $q$ follows an Antoniak distribution [30] as

$$q | N', U' \sim \text{Antoniak}(N', U').$$  \hspace{1cm} (68)

The Antoniak distribution (also called the Chinese Restaurant Table distribution [31]) 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 $q \sim \sum_{w=1}^{N'} \text{Ber}(U'/(U' + w - 1))$. Given $q$ and assuming gamma prior as $U' \sim \text{Gam}(e_0, f_0)$, the posterior for $U'$ is given by

$$P(U' | N', q) \propto P(q | N', U')P(U') \propto U'^q \times \frac{1}{\Gamma(N)} U'^{e_0-1}e^{-f_0}.$$  \hspace{1cm} (69)

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

$$U' | N', q \sim \text{Gam}(e_0 + q, f_0).$$  \hspace{1cm} (70)

### 4.3.2 Sampling Hyperparameters

In this section, we show that posterior samplers for all hyperparameters (i.e., $\gamma^1$, $\gamma^2$, $\beta$, $\alpha^1$, $\alpha^2$, and $a$) can be derived using the data augmentation techniques we introduced in Sect. 4.3.1.

Since a beta function is equivalently rewritten as $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, Eq. (28) can be equivalently rewritten as follows:

$$P(R | z^1, z^2, r) = \frac{\Gamma(2\beta)}{\Gamma(\beta)^2} \prod_{k} \left\{ \frac{\Gamma(m_{k,i} + \bar{m}_{k,i} + \beta)}{\Gamma(\beta)} \times \frac{\Gamma(m_{k,i} + \beta)}{\Gamma(\beta)} \right\}.$$  \hspace{1cm} (71)

Therefore, by expanding Eq. (71) using Eqs. (61) and (62), we can obtain a joint distribution over $\beta$ and several auxiliary variables, where posterior samplers for the auxiliary variables are derived as follows:

$$p_{\beta} \sim \text{Beta}(2\beta, m_{\beta} + \bar{m}_{\beta} + m_{\beta} + \bar{m}_{\beta}),$$  \hspace{1cm} (72)

$$q_{\beta} \sim \text{Antoniak}(m_{\beta} + \bar{m}_{\beta}, \beta),$$  \hspace{1cm} (73)

$$\bar{q}_{\beta} \sim \text{Antoniak}(m_{\beta} + \bar{m}_{\beta}, \beta),$$  \hspace{1cm} (74)

$$p_{k,i} \sim \text{Beta}(2\beta, m_{k,i} + \bar{m}_{k,i} + m_{k,i} + \bar{m}_{k,i}),$$  \hspace{1cm} (75)

$$q_{k,i} \sim \text{Antoniak}(m_{k,i} + \bar{m}_{k,i}, \beta),$$  \hspace{1cm} (76)

$$\bar{q}_{k,i} \sim \text{Antoniak}(m_{k,i} + \bar{m}_{k,i}, \beta).$$  \hspace{1cm} (77)

Consequently, assuming the prior as $\beta \sim \text{Gam}(e_0, f_0)$, $\beta$ can be updated as

$$\beta | - \sim \text{Gam}(e_0 + q_{\beta} + \bar{q}_{\beta} + \sum_k \sum_i (q_{k,i} + \bar{q}_{k,i}),$$

$$f_0 - 2 \ln p_{\beta} - \sum_k \sum_i \ln p_{k,i}),$$  \hspace{1cm} (78)

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

For $a$, $\alpha^1$, $\alpha^2$, $\gamma^1$, and $\gamma^2$, posterior samplers can be straightforwardly derived by expanding Eqs. (35), (36), (37), (40), and (41) using Eqs. (61) and (62). Consequently, these hyperparameters can be updated as follows:

$$p_{\alpha^1} \sim \text{Beta} \left( \sum_s a_{s,t,u}, \sum_s n_{s,t,u} \right),$$  \hspace{1cm} (79)

$$q_{s,t,u} \sim \text{Antoniak}(n_{s,t,u}, a_{s,t,u}),$$  \hspace{1cm} (80)

$$a_{s,t,u} | - \sim \text{Gam}(e_0 + q_{s,t,u}, f_0 - \ln p_{s,t,u}),$$  \hspace{1cm} (81)

$$p \sim \text{Beta} \left( \sum_s \alpha^1_s, N^2 \right),$$  \hspace{1cm} (82)

$$q_{i,s} \sim \text{Antoniak}(n_{i,s}, \alpha^1_s),$$  \hspace{1cm} (83)

$$\alpha^1_s | - \sim \text{Gam} \left( e_0 + \sum_i q_{i,s}, f_0 - I \ln p \right),$$  \hspace{1cm} (84)
\[ p \sim \text{Beta}(\gamma^1, N^1), \gamma^1 | - \sim \text{Gam}(e_0 + K, f_0 - \ln p). \] (85)

Note that posterior samplers for \( \alpha^2 \) and \( \gamma^2 \) are omitted because these can be sampled in the same way as \( \alpha^1 \) and \( \gamma^1 \), respectively.

### 4.4 Pseudocode for Performing Posterior Inference

Using the posterior samplers derived in Sects. 4.2 and 4.3, 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.

### 4.5 Computational Efficiency

Here, we briefly discuss the computational cost of the MLIRM and related models. As evident from the form of Eq. (42), posterior update for a cluster assignment \( z_i^1 \) requires \( O(KL) \) computation. Therefore, updating \( z_i^1 \) and \( z_i^2 \) requires \( O((I + J)KL) \) computation. Additionally, the MLIRM requires \( O(IJ) \) computation to update auxiliary variables \( r_i^1, r_i^2, \) and \( r \). 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 [32]. 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 [25], [33], [34] is one of the promising directions for future work.

### 5. Experiments

In this section, we present the experimental results using a toy synthetic dataset and several real-world datasets, i.e., “Animal” [35], “Enron” [36], and “MovieLens” [37]. In all experiments, we fit the hyperparameters of the MLIRM using the samplers we derived in Sect. 4.2. 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.

#### 5.1 Synthetic Data

First, we explored the ability of the proposed MLIRM to recover the underlying cluster structure using synthetic data. Figure 6 (a) shows the hand-constructed synthetic data used in this experiment. As can be seen in Fig. 6 (a), there are...
many biased objects with extremely many links or few links. Figures 6(b)–6(k) show the clustering results obtained by the MLIRM and several of the related models that were reviewed in Sect. 3.3.

As shown in Fig. 6(b), 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. 6(c), 6(d), and 6(e)) 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. 6(f) and 6(g)), either the clustering layer or the background layer is selected in an object-wise manner. Therefore, the SIRM cannot consider that an observation is a relevant entry. For quantitative comparison, three measurements were used; i.e., train error (0-1 loss), test log likelihood, and number of obtained cluster blocks. 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. The train error (∈ [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 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 1 lists the computed measures. In the case of every dataset for all measurements, except for the number of clusters, the MLIRM significantly outperformed the other models. For the “MovieLens” dataset, the MLIRM and RDIRM-prod obtained nearly identical values for the number of clusters 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.

5.2 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 rating; 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 $j$.

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 ($\epsilon \in [0, 1]$) 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 1 lists the computed measures. In the case of every dataset for all measurements, except for the number of cluster blocks, the MLIRM significantly outperformed the other models. For the “MovieLens” dataset, the MLIRM and RDIRM-prod obtained nearly identical values for the number of cluster blocks 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 7 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 $\phi$, we illustrate the most probable layer obtained for each dataset in Fig. 9. As can be seen in Fig. 9, 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. 7 (d) and 7 (e), 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:

\[
S^1_u(i) := \frac{\sum_{y \in T^2} I(r_{i,y} = u)}{N^2}, \quad S^2_u(j) := \frac{\sum_{x \in T^1} I(r_{x,j} = u)}{N^1}.
\]

(86)

Table 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 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. 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 8 depicts the detailed clustering results for the “Animal” dataset.

Table 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          | \( S^2_u(j) \) | \( z^2 \) | feature          | \( S^2_u(j) \) |
|------------------|----------------|---------|------------------|----------------|
| meat             | 0.98           | 15      | newworld         | 0.74           |
| hunter           | 0.86           | 15      | chewteeth        | 0.72           |
| meatteeth        | 0.84           | 15      | black            | 0.60           |
| fierce           | 0.62           | 15      | oldworld         | 0.57           |
| fast             | 1.00           | 21      | solitary         | 0.40           |
| tail             | 0.98           | 21      | white            | 0.38           |
| agility          | 0.96           | 21      | feature          | \( S^2_u(j) \) |
| swims            | 0.98           | 22      | orange           | 0.80           |
| coastal          | 0.98           | 22      | yellow           | 0.80           |
| strainteeth      | 0.98           | 22      | skimmer          | 0.78           |
| claws            | 1.00           | 24      | bush             | 0.76           |
| paws             | 0.84           | 24      | stripes          | 0.76           |
| nocturnal        | 0.76           | 24      | desert           | 0.72           |

Fig. 7  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 \( S^1_u(i) \) and \( S^2_u(j) \)).

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 Figs. 7(b), 7(f), 7(g), and 7(h), 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. 7(b)). 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. 7(g)). This employee was a key person in the 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.

6. Conclusions

In this study, we have addressed the problem of noisy relational data in which the underlying cluster structure is blurred by biased objects. To overcome this problem, we have 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 co-cluster 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 study. 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 compare 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 [25], [33] or gradient-based stochastic methods [34]. 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 [6], [12], [38] or multiple membership models [7], [8], [39].

References

[1] M.M. Shafiei and E.E. Milios, “Latent Dirichlet co-clustering,” Proc. ICDM, pp.542–551, 2006.
[2] Z. Xu, V. Tresp, K. Yu, and H. Kriegel, “Infinite hidden relational models,” Proc. UAI, pp.544–551, 2006.
[3] H. Shan and A. Banerjee, “Bayesian co-clustering,” Proc. ICDM, pp.530–539, 2008.
[4] K. Nowicki and T.A.B. Snijders, “Estimation and prediction for stochastic blockstructures,” J. Am. Stat. Assoc., vol.96, no.455, pp.1077–1087, 2001.
[5] C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda, “Learning systems of concepts with an infinite relational model,” Proc. AAAI, pp.381–388, 2006.
[6] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing, “Mixed membership stochastic blockmodels,” J. Mach. Learn. Res., vol.9, pp.1981–2014, 2008.
[7] M. Mørup, M.N. Schmidt, and L.K. Hansen, “Infinite multiple membership relational modeling for complex networks,” Proc. MLSP, pp.1–6, 2011.
[8] K. Palla, D.A. Knowles, and Z. Ghahramani, “An infinite latent attribute model for network data,” Proc. ICML, pp.1607–1614, 2012.
[9] D.M. Roy, C. Kemp, V.K. Mansinghka, and J.B. Tenenbaum, “Learning annotated hierarchies from relational data,” Proc. NIPS, pp.1185–1192, 2006.
[10] M. Nakano, K. Ishiguro, A. Kimura, T. Yamada, and N. Ueda, “Rectangular tiling process,” Proc. ICML, pp.361–369, 2014.
[11] K. Ishiguro, I. Sato, M. Nakano, A. Kimura, and N. Ueda, “Infinite plaid models for infinite bi-clustering,” Proc. AAAI, pp.1701–1708, 2016.
[12] W. Fu, L. Song, and E.P. Xing, “Dynamic mixed membership block-model for evolving networks,” Proc. ICML, pp.329–336, 2009.
[13] K. Ishiguro, T. Iwata, N. Ueda, and J.B. Tenenbaum, “Dynamic infinite relational model for time-varying relational data analysis,” Proc. NIPS, pp.919–927, 2010.
[14] Q. Ho, A.P. Parikh, L. Song, and E.P. Xing, “Multiscale community blockmodel for network exploration,” Proc. AISTATS, pp.333–341, 2011.
[15] I. Ohama, T. Kida, and H. Arimura, “Multi-layered framework for modeling relationships between biased objects,” Proc. SDM, pp.819–827, 2015.
[16] D. Blackwell and J.B. MacQueen, “Ferguson distributions via Polya
urn schemes,” The Annals of Statistics, vol.1, no.2, pp.353–355, 1973.
[17] D. Aldous, “Exchangeability and related topics,” École d’Été de Probabilités de Saint-Flour XIII, Lecture Notes in Mathematics, vol.1117, pp.1–198, Springer, Berlin, Heidelberg, 1985.
[18] J.S. Liu, “The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem,” J. Am. Stat. Assoc., vol.89, no.427, pp.958–966, 1994.
[19] T.S. Ferguson, “A Bayesian analysis of some nonparametric problems,” The Annals of Statistics, vol.1, no.2, pp.209–230, 1973.
[20] P.D. Hoff, “Subset clustering of binary sequences, with an application to genomic abnormality data,” Biometrics, vol.61, no.4, pp.1027–1036, 2005.
[21] P.D. Hoff, “Model-based subspace clustering,” Bayesian Analysis, vol.1, no.2, pp.321–344, 2006.

K. Ishiguro, N. Ueda, and H. Sawada, “Subset infinite relational models,” Proc. AISTATS, pp.547–555, 2012.
[22] I. Ohama, H. Iida, T. Kida, and H. Arimura, “An extension of the infinite relational model incorporating interaction between objects,” Proc. PAKDD, Lecture Notes in Computer Science, vol.7819, pp.147–159, Springer, Berlin, Heidelberg, 2013.
[23] I. Ohama, H. Iida, T. Kida, and H. Arimura, “The relevance dependent infinite relational model for discovering co-cluster structure from relationships with structured noise,” IEICE Trans. Inf. & Syst., vol.E99-D, no.4, pp.1139–1152, April 2016.
[24] Y.W. Teh, D. Newman, and M. Welling, “A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation,” Proc. NIPS, pp.1353–1360, 2006.
[25] M.D. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” J. Am. Stat. Assoc., vol.90, no.430, pp.577–588, 1995.
[26] M. Zhou, “Infinite edge partition models for overlapping community detection and link prediction,” Proc. AISTATS, pp.1135–1143, 2015.
[27] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical Dirichlet processes,” J. Am. Stat. Assoc., vol.101, no.476, pp.1566–1581, 2006.
[28] D. Newman, A.U. Asuncion, P. Smyth, and M. Welling, “Distributed algorithms for topic models,” J. Mach. Learn. Res., vol.10, pp.1801–1829, 2009.
[29] C.E. Antoniak, “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” The Annals of Statistics, vol.2, no.6, pp.1152–1174, 1974.
[30] M. Zhou and L. Carin, “Negative binomial process count and mixture modeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.2, pp.307–320, 2015.
[31] I. Ohama, T. Kida, and H. Arimura, “Discovering relevance-dependent bicluster structure from relational data,” Proc. IJCAI, pp.2578–2584, 2017.
[32] K. Ishiguro, I. Sato, and N. Ueda, “Averaged collapsed variational Bayes inference,” J. Mach. Learn. Res., vol.18, no.1, pp.1–29, 2017.
[33] Y. Cong, B. Chen, H. Liu, and M. Zhou, “Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC,” Proc. ICML, pp.864–873, 2017.
[34] D.N. Osherson, J. Stern, O. Wilkie, M. Stob, and E.E. Smith, “Default probability,” Cognitive Science, vol.15, no.2, pp.251–269, 1991.
[35] B. Klimat and Y. Yang, “The Enron corpus: A new dataset for email classification research,” Proc. ECML, Lecture Notes in Computer Science, vol.3201, pp.217–226, Springer, Berlin, Heidelberg, 2004.
[36] M. Zhou, “Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling,” Proc. NIPS, pp.3455–3463, 2014.
[37] I. Ohama, I. Sato, T. Kida, and H. Arimura, “On the model shrinkage effect of gamma process edge partition models,” Proc. NIPS, pp.396–404, 2017.

Iku Ohama received his B.S. in Engineering from Kyoto University in 2004. He received his M.S. in Informatics from Kyoto University in 2006. In 2006, he joined the Corporate R&D Division, Matsushita Electric Industrial (now Panasonic) Corporation, Osaka Japan. From 2006 to 2008, he was engaged in research and development of the software platform for consumer electronics. Since 2009, he has been engaged in research and development of machine learning and data mining technologies for knowledge discovery from data in business fields. His research interests include statistical machine learning, latent variable models, Bayesian nonparametrics, matrix factorization, and its related algorithms in these fields.

Takuya Kida received the B.S. degree in Physics, the M.S. and Dr. (Information Science) degree all from Kyushu University in 1997, 1999, and 2001, respectively. He was a full-time lecturer of Kyushu University Library from October 2001 to March 2004. He is currently an associate professor of the division of computer science, Graduate School of Information Science and Technology, Hokkaido University, since April 2004. His research interests include text algorithms and data structures, information retrieval, and data compression. He is a member of IEICE, IPSJ, and DBSJ.

Hiroki Arimura received the B.S. degree in 1988 in Physics, the M.S. and Dr. (Information Science) degree all from Kyushu University from 1990 to 1996. He has been engaged in research and data mining, computational learning theory, information retrieval, artificial intelligence, and algorithm theory. He is a member of ACM, IEICE, IPSJ, and JSAI.