Discovering Relevance-Dependent Bicluster Structure from Relational Data

Iku Ohama[†‡], Takuya Kida[‡], and Hiroki Arimura[‡]
[†]Panasonic Corporation, Japan
[‡]Graduate School of Information Science and Technology, Hokkaido University, Japan
ohama.iku@jp.panasonic.com, {kida,arim}@ist.hokudai.ac.jp

Abstract

In this paper, we propose a statistical model for *relevance-dependent biclustering* to analyze relational data. The proposed model factorizes relational data into bicluster structure with two features: (1) each object in a cluster has a *relevance* value, which indicates how strongly the object relates to the cluster and (2) all clusters are related to at least one dense block. These features simplify the task of understanding the meaning of each cluster because only a few highly relevant objects need to be inspected. We introduced the Relevance-Dependent Bernoulli Distribution (R-BD) as a prior for relevance-dependent binary matrices and proposed the novel Relevance-Dependent Infinite Biclustering (R-IB) model, which automatically estimates the number of clusters. Posterior inference can be performed efficiently using a collapsed Gibbs sampler because the parameters of the R-IB model can be fully marginalized out. 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.

1 Introduction

Relational data encoding pairwise relationships between I and J objects appears in many fields. *Biclustering* is one of the most popular techniques to extract useful insights from relational data. It abstracts the given data matrix to a low-dimensional block structure by simultaneously clustering both the row and column objects. For example, biclustering of point-of-sale (POS) data can be used to elucidate bipartite relationships between particular customers and particular items that sell well. In this paper, we focused on the most typical type of relational data, i.e., sets of observations represented as a binary matrix.

Figure 1a shows standard biclustering, which is the basic concept assumed in many existing biclustering models [Nowicki and Snijders, 2001; Kemp *et al.*, 2006; Airoldi *et al.*, 2008; Xu *et al.*, 2006]. In particular, the Infinite Relational Model (IRM) [Kemp *et al.*, 2006] is one of the most widely used biclustering models, and it can automatically estimate the number of clusters. In such standard biclustering models, it is assumed that the given relational data is abstracted by a block structure in which each block has uniform density. However, this assumption has problems when we analyze relational data. We typically analyze an obtained cluster from the objects constituting the cluster. Thus, the assumption requires that we inspect all objects in the rows and columns of each block to understand what the block means. Furthermore, biclustering models with this assumption often extract many meaningless clusters that comprise objects with few links [Ishiguro *et al.*, 2012; 2016]. Thus, extracting insights from standard biclustering results is often very time-consuming.

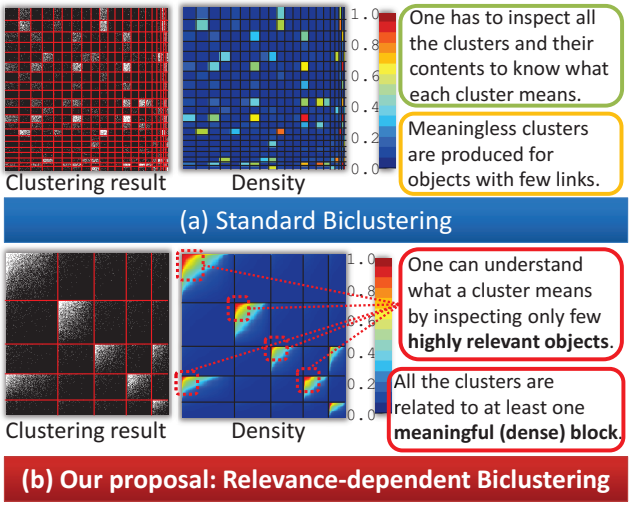


Figure 1: Diagrams of (a) standard biclustering and (b) relevance-dependent biclustering.

In this paper, we propose relevance-dependent biclustering (Fig. 1b) to solve the drawbacks of standard biclustering. We assume that each row and column object has a hidden variable indicating a *relevance* value that determines how strongly the object relates to the cluster. A large relevance value means that the corresponding object strongly relates to the cluster. Thus, such highly relevant objects ought to be inspected to aid in understanding the meaning of the clusters. In other

words, a small relevance value indicates that the corresponding object is non-informative. The advantages of relevance-dependent biclustering are as follows:

- The meaning of obtained clusters can be understood by inspecting only a few highly relevant objects.
- All obtained clusters are interpretable because they are related to at least one meaningful (dense) block.

Consequently, the meaning of the biclustering results can be understood easily without cumbersome manual effort.

Now we provide a motivating example for the proposed biclustering. When analyzing the shopping behavior of customers on an e-commerce 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 a bicluster structure in such a situation, it is natural to assume additional latent factors (i.e., relevance values) that affect the link probabilities of objects regardless of their cluster membership. Although the necessity of considering clusters with uneven density has been discussed in social community analysis [Araujo *et al.*, 2014], this problem has not yet been studied directly. Another challenge is to develop an efficient inference algorithm for this problem, as simultaneous estimation of bicluster structure and relevance is a chicken-and-egg problem.

Contributions: We have four contributions in this paper.

1. For modeling relevance-dependent binary matrices, we propose *Relevance-Dependent Bernoulli Distribution (R-BD)*, which is parameterized by a typical link strength common to all entries in the matrix and relevance parameters for each row and column object. We also propose conjugate priors for R-BD. Thus, all R-BD parameters can be marginalized out.
2. By incorporating R-BD, we propose a novel biclustering model called the *Relevance-Dependent Infinite Biclustering (R-IB)* model, which can extract relevance-dependent bicluster structure from relational data with an unknown number of clusters.
3. The R-IB has an efficient inference algorithm because all model parameters can be marginalized out and do not need to be estimated explicitly.
4. In experiments, we quantitatively confirm that the R-IB can capture more essential bicluster structure with better computational efficiency than conventional models. In addition, we show the ability of the R-IB to derive an interpretable bicluster structure that facilitates the extraction of insights from real-world relational data.

The rest of this paper is organized as follows. We describe existing models in Sec. 2. We propose our novel model and its inference algorithm in Sec. 3. We present experimental results in Sec. 4. Finally, we conclude the paper in Sec. 5.

2 Existing Models

First, we review the IRM [Kemp *et al.*, 2006] as a baseline standard biclustering model and then review existing IRM extensions that consider the relevance of objects.

2.1 Infinite Relational Model

Let $\boldsymbol{R}$ be $I \times J$ relational data between I row objects and J column objects. $R_{i,j} = 1(0)$ indicates the existence of a link (non-link) between the i-th row and the j-th column. The IRM has latent variables $z_{1,i} \in \{1, \cdots, K\}$ and $z_{2,j} \in \{1, \cdots, L\}$ that indicate the cluster assignments for I and J objects, respectively. In addition, let $\boldsymbol{\eta} \in [0,1]^{K \times L}$ be a link probability matrix between K row clusters and L column clusters. Then, the link between the i-th row object and the j-th column object is drawn as follows:

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

In the IRM, the Chinese Restaurant Process (CRP) [Blackwell and MacQueen, 1973; Aldous, 1985] is used as the prior for cluster assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. Thus, the IRM estimates the number of clusters automatically from the observed data. Thanks to the conjugacy between beta and Bernoulli distributions, $\boldsymbol{\eta}$ can be marginalized out. Consequently, the posterior inference for the IRM can be performed efficiently using collapsed methods [Liu, 1994; Teh *et al.*, 2006b].

As the IRM is widely used, many extensions have been proposed to extract more advanced cluster structures, such as the mixed membership structure [Airoldi *et al.*, 2008], the hierarchical structure [Roy *et al.*, 2006], the multiple membership structure [Mørup *et al.*, 2011; Palla *et al.*, 2012], and other extensions [Nakano *et al.*, 2014; Ho *et al.*, 2011; Ishiguro *et al.*, 2010; Fu *et al.*, 2009]. However, none of them provides relevance information, because these models also assume that the given data exactly follows the density defined by only the cluster structure. As these advanced models can be considered instances of standard biclustering models, we focus on the IRM as the baseline model.

2.2 Existing Relevance Modeling

Few studies consider relevance dependency in biclustering. The Relevance Dependent Infinite Relational Model (RDIRM) [Ohama *et al.*, 2013; 2016] 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:

$$\begin{gathered} R_{i,j} \sim \text{Bernoulli}(r_{i,j} \times \eta_{z_{1,i},z_{2,j}} + (1 - r_{i,j}) \times \eta_0), \\ r_{i,j} = F(r_{1,i\to j}, r_{2,j\to i}), \\ r_{1,i\to j} \sim \text{Bernoulli}(\rho_{1,i}), \;\; r_{2,j\to i} \sim \text{Bernoulli}(\rho_{2,j}), \end{gathered} \tag{2}$$

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

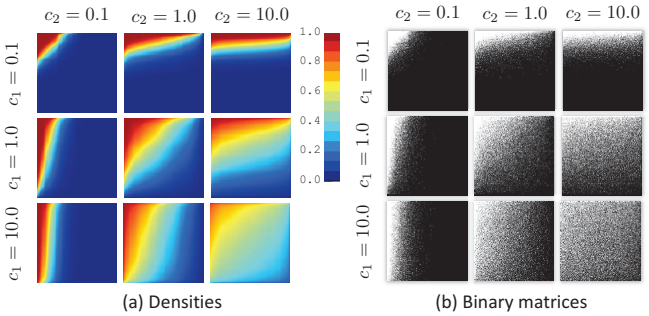


Figure 2: Outcomes drawn from the R-BD ($\lambda = 0.6931$) with different Dirichlet parameters. (a) shows probability densities and (b) shows binary matrices drawn from corresponding density, where black corresponds to 0 and white to 1.

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.* [2015] 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.

3 Proposed Model

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.

3.1 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, $\theta_{1,i}$, and $\theta_{j,2}$. 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 $\theta_{1,i}$ and $\theta_{2,j}$ (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 $\theta_{i,1}\theta_{j,2}\lambda$. Finally, to obtain a binary random variable, we define Relevance-dependent Bernoulli distribution (R-BD) as follows:

$$x_{i,j} \sim \text{Bernoulli}\left(1 - e^{-\theta_{1,i}\theta_{2,j}\lambda}\right), \tag{3}$$

where the function $f(s) = 1 - e^{-s}$ is the Bernoulli-Poisson (BerPo) link function [Zhou, 2015] 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, $\boldsymbol{\theta_1}$, and $\boldsymbol{\theta_2}$) can be marginalized out. Following the property of BerPo link, Eq. (3) can be equivalently rewritten by truncating a Poisson random variable $x^*_{i,j}$ as

$$x_{i,j} = \mathbb{I}(x^*_{i,j} \geq 1),\ x^*_{i,j} \sim \text{Poisson}(\theta_{1,i}\theta_{2,j}\lambda), \tag{4}$$

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:

$$x^* \,|\, x, \lambda \sim \begin{cases} \delta(0), & \text{if } x = 0 \\ \text{ZTP}(\lambda). & \text{if } x = 1 \end{cases} \tag{5}$$

Note that $\delta(0)$ is a point mass at zero and ZTP$(\cdot)$ denotes a *zero-truncated Poisson* distribution [Geyer, 2007]. 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], $\{\theta_{1,i}\}_{i=1}^{I}/I \sim \text{Dirichlet}(c_1)$, and $\{\theta_{2,j}\}_{j=1}^{J}/J \sim \text{Dirichlet}(c_2)$, we obtain a closed-form marginal likelihood for auxiliary counts $\boldsymbol{x}^*$ as follows:

$$P(\boldsymbol{x}^*) = \frac{1}{\prod_{i,j} x^*_{i,j}!} \prod_i \frac{\Gamma(c_1 + M_{i,\cdot})}{\Gamma(c_1)} \prod_j \frac{\Gamma(c_2 + M_{\cdot,j})}{\Gamma(c_2)} \times \frac{I^M \Gamma(Ic_1)}{\Gamma(Ic_1 + M)} \frac{J^M \Gamma(Jc_2)}{\Gamma(Jc_2 + M)} \frac{G(a + M, b + IJ)}{G(a, b)}, \tag{6}$$

where $M_{i,\cdot} = \sum_j x^*_{i,j}$, $M_{\cdot,j} = \sum_i x^*_{i,j}$, and $M = \sum_{i,j} x^*_{i,j}$. 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 2 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 value of a given binary matrix.

3.2 Relevance-Dependent Infinite Biclustering

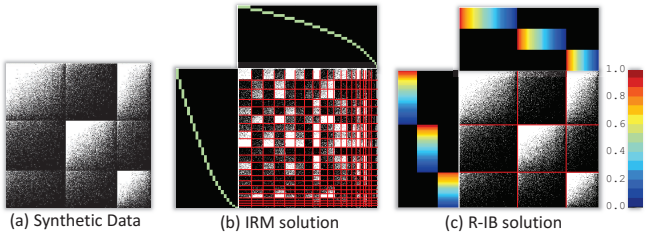
Here, we describe the proposed R-IB model. CRP(γ) denotes a CRP with concentration parameter γ. The full description of the R-IB model, incorporating R-BD to the observation model of the IRM, is as follows:

$$\begin{aligned} R_{i,j} \,|\, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\lambda} &\sim \text{Bernoulli}\left(1 - e^{-\theta_{1,i}\theta_{2,j}\lambda_{z_{1,i},z_{2,j}}}\right), \\ \lambda_{k,l} \,|\, a, b &\sim \text{Gamma}(a, b), \\ \boldsymbol{\theta}_{1,k}/n_{1,k} \,|\, c_1, \boldsymbol{z}_1 &\sim \text{Dirichlet}(\overbrace{c_1, \cdots, c_1}^{n_{1,k}}), \\ \boldsymbol{\theta}_{2,l}/n_{2,l} \,|\, c_2, \boldsymbol{z}_2 &\sim \text{Dirichlet}(\overbrace{c_2, \cdots, c_2}^{n_{2,l}}), \\ z_{1,i} \,|\, \gamma_1 \sim \text{CRP}(\gamma_1),\ z_{2,j} &\,|\, \gamma_2 \sim \text{CRP}(\gamma_2), \end{aligned} \tag{7}$$

where $n_{1,k}$ ($n_{2,l}$) is the number of row (column) objects assigned to cluster k (l). Note that $\boldsymbol{\theta}_{1,k}$ ($\boldsymbol{\theta}_{2,l}$) is a set of relevance parameters $\theta_{1,i}$ ($\theta_{2,j}$), 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 $\boldsymbol{\lambda}$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ can be marginalized out. The marginal

[1] 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.



(a) Synthetic Data (b) IRM solution (c) R-IB solution

Figure 3: 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 θ is transformed into a probability in $[0, 1]$ as $1 - e^{-\log(2)\times\theta}$.

likelihood for $\boldsymbol{R}^*$, given $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, is then given as

$$
\begin{aligned}
&P(\boldsymbol{R}^* \mid \boldsymbol{z}_1, \boldsymbol{z}_2) \\
&= \frac{1}{\prod_{i,j} R^*_{i,j}!} \prod_i \frac{\Gamma(c_1 + M_{i,\cdot})}{\Gamma(c_1)} \prod_j \frac{\Gamma(c_2 + M_{\cdot,j})}{\Gamma(c_2)} \\
&\quad \times \prod_k \frac{{n_{1,k}}^{M_{k,\cdot}} \Gamma(n_{1,k} c_1)}{\Gamma(n_{1,k} c_1 + M_{k,\cdot})} \prod_l \frac{{n_{2,l}}^{M_{\cdot,l}} \Gamma(n_{2,l} c_2)}{\Gamma(n_{2,l} c_2 + M_{\cdot,l})} \\
&\quad \times \prod_k \prod_l \frac{G(a + M_{k,l}, b + n_{1,k} n_{2,l})}{G(a, b)},
\end{aligned} \tag{8}
$$

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

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

3.3 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 (8) and the likelihood for the CRP, given $\boldsymbol{R}^*$, the posterior probability that the i-th object is assigned to cluster k^* is given by

$$
\begin{aligned}
&P(z_{1,i} = k^* \mid -) \\
&\propto \begin{cases}
n^{-i}_{1,k^*} \times \frac{\left(n^{+i}_{1,k^*}\right)^{M^{+i}_{k^*,\cdot}} \Gamma(n^{+i}_{1,k^*} c_1) \Gamma(n^{-i}_{1,k^*} c_1 + M^{-i}_{k^*,\cdot})}{\left(n^{-i}_{1,k^*}\right)^{M^{-i}_{k^*,\cdot}} \Gamma(n^{-i}_{1,k^*} c_1) \Gamma(n^{+i}_{1,k^*} c_1 + M^{+i}_{k^*,\cdot})} \\
\quad \times \prod_l \frac{G(a + M^{+i}_{k^*,l}, b + n^{+i}_{1,k^*} n_{2,l})}{G(a + M^{-i}_{k^*,l}, b + n^{-i}_{1,k^*} n_{2,l})}, & \text{if } n^{-i}_{1,k^*} > 0 \\
\gamma_1 \times \frac{\Gamma(c_1)}{\Gamma(c_1 + M_{i,\cdot})} \\
\quad \times \prod_l \frac{G(a + M^{+i}_{k^*,l}, b + n^{+i}_{1,k^*} n_{2,l})}{G(a, b)}, & \text{if } n^{-i}_{1,k^*} = 0
\end{cases}
\end{aligned} \tag{9}
$$

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

$$
R^*_{i,j} \sim \begin{cases} \delta(0), & \text{if } R_{i,j} = 0 \\ \text{ZTP}(\theta_{1,i} \theta_{2,j} \lambda_{z_{1,i}, z_{2,j}}). & \text{if } R_{i,j} = 1 \end{cases} \tag{10}
$$

Note that explicit samples for $\boldsymbol{\lambda}$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ are only required during the sampling of $\boldsymbol{R}^*$, and are drawn as follows: $\lambda_{k,l} \mid - \sim \text{Gamma}(a + M_{k,l}, b + n_{1,k} n_{2,l})$, $\boldsymbol{\theta}_{1,k} / n_{1,k} \mid - \sim \text{Dirichlet}(c_1 + \boldsymbol{M}_{1,k})$, and $\boldsymbol{\theta}_{2,l} / n_{2,l} \mid - \sim \text{Dirichlet}(c_2 + \boldsymbol{M}_{2,k})$, 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.

The hyperparameters for the R-IB (i.e., γ_1, γ_2, c_1, c_2, a, and b) can also be sampled assuming gamma priors. Thanks to the conjugacy between gamma distributions, posterior sampling for the rate parameter b is straightforward. For the remaining hyperparameters, posterior sampling is performed using *data augmentation* techniques [Escobar and West, 1994; Zhou, 2015; Teh *et al.*, 2006a; Newman *et al.*, 2009] (omitted here for brevity).

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. 4.2).
- To show the usefulness of relevance-dependent biclustering results obtained by the R-IB in understanding the meaning of each cluster (Sec. 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)).

4.1 Datasets

The first dataset was the Animal [Osherson *et al.*, 1991] 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 [Klimat and Yang, 2004] 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

Table 1: Computed AUC-PR on real-world datasets. Best results are highlighted in bold. Parenthesized numbers indicate standard deviations.

	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)

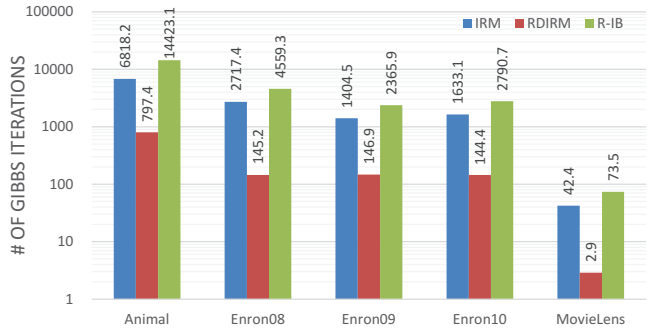


Figure 4: Average number of Gibbs iterations per five minutes in logarithmic scale. All the models were implemented in JAVA and run on a PC with an Intel® Xeon® 2.7 GHz CPU.

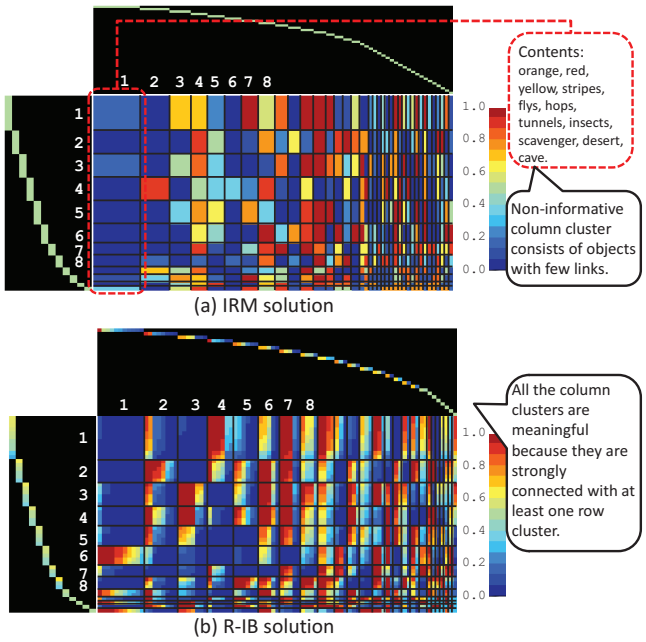


Figure 5: Clustering results on the Animal dataset. The central color map denotes estimated link probabilities.

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 [MOV, as of 2003] 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.

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 the Precision-Recall curve (AUC-PR) [Davis and Goadrich, 2006]. We compared three biclustering models: the IRM, the RDIRM, and the R-IB. We ran 4000 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 1 lists the results. As can be seen, R-IB significantly outperformed the RDIRM with all datasets. 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 superiority of our R-IB over conventional models in link prediction accuracy.

The average number of Gibbs iterations within 5 min was used as the metric to evaluate the computational efficiency of the different models. As shown in Fig. 4, 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. 3.3.

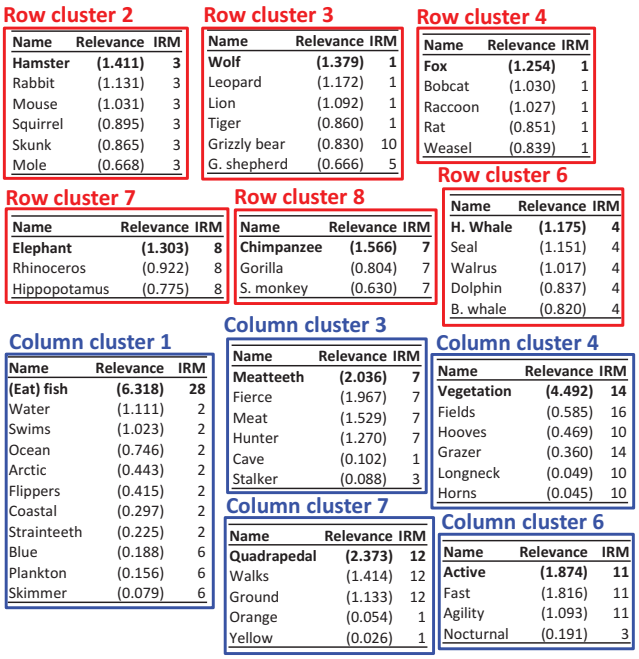
These quantitative results confirmed that the R-IB can extract more essential bicluster structures with better computational efficiency than conventional models.

4.3 Qualitative Comparison

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

Figure 5 shows the clustering results on the Animal dataset. As can be seen from Fig. 5a 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. 5a). In contrast, as Fig. 5b 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. 6 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”



Row cluster 2

Name	Relevance	IRM
Hamster	**(1.411)**	**3**
Rabbit	(1.131)	3
Mouse	(1.031)	3
Squirrel	(0.895)	3
Skunk	(0.865)	3
Mole	(0.668)	3

Row cluster 3

Name	Relevance	IRM
Wolf	**(1.379)**	**1**
Leopard	(1.172)	1
Lion	(1.092)	1
Tiger	(0.860)	1
Grizzly bear	(0.830)	10
G. shepherd	(0.666)	5

Row cluster 4

Name	Relevance	IRM
Fox	**(1.254)**	**1**
Bobcat	(1.030)	1
Raccoon	(1.027)	1
Rat	(0.851)	1
Weasel	(0.839)	1

Row cluster 6

Name	Relevance	IRM
H. Whale	**(1.175)**	**4**
Seal	(1.151)	4
Walrus	(1.017)	4
Dolphin	(0.837)	4
B. whale	(0.820)	4

Row cluster 7

Name	Relevance	IRM
Elephant	**(1.303)**	**8**
Rhinoceros	(0.922)	8
Hippopotamus	(0.775)	8

Row cluster 8

Name	Relevance	IRM
Chimpanzee	**(1.566)**	**7**
Gorilla	(0.804)	7
S. monkey	(0.630)	7

Column cluster 1

Name	Relevance	IRM
(Eat) fish	**(6.318)**	**28**
Water	(1.111)	2
Swims	(1.023)	2
Ocean	(0.746)	2
Arctic	(0.443)	2
Flippers	(0.415)	2
Coastal	(0.297)	2
Strainteeth	(0.225)	2
Blue	(0.188)	6
Plankton	(0.156)	6
Skimmer	(0.079)	6

Column cluster 3

Name	Relevance	IRM
Meatteeth	**(2.036)**	**7**
Fierce	(1.967)	7
Meat	(1.529)	7
Hunter	(1.270)	7
Cave	(0.102)	1
Stalker	(0.088)	3

Column cluster 4

Name	Relevance	IRM
Vegetation	**(4.492)**	**14**
Fields	(0.585)	16
Hooves	(0.469)	10
Grazer	(0.360)	14
Longneck	(0.049)	10
Horns	(0.045)	10

Column cluster 7

Name	Relevance	IRM
Quadrapedal	**(2.373)**	**12**
Walks	(1.414)	12
Ground	(1.133)	12
Orange	(0.054)	1
Yellow	(0.026)	1

Column cluster 6

Name	Relevance	IRM
Active	**(1.874)**	**11**
Fast	(1.816)	11
Agility	(1.093)	11
Nocturnal	(0.191)	3

Figure 6: 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.

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. 6) 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. 6). 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 characteristics. 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 7 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. 7a). Although the IRM also produced a comparatively large cluster block comprising many moderately strongly linked objects (block B in Fig. 7a), it was difficult to understand the meaning of the block. In contrast, R-IB produced a more interpretable bicluster structure (Fig. 7b). 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. 7b, 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. 7b were interpreted as "mails between cash analysts" and "reports from employees to the chief financial officer," respectively.

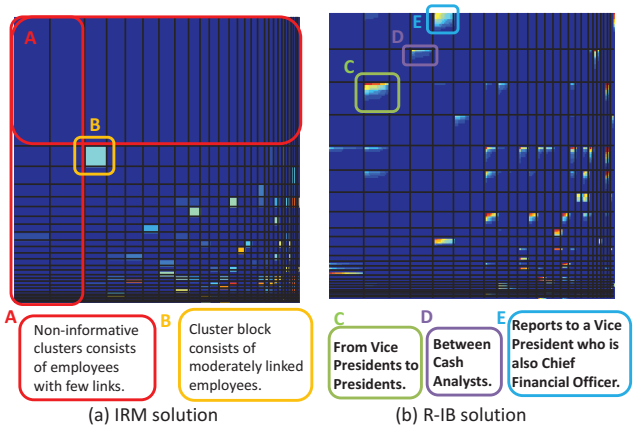


Figure 7: Clustering results on the Enron09 dataset.

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 Conclusions

In this paper, 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. 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. In the future, we intend to extend this to be able to discover more advanced structures, such as those with mixed membership or multiple membership assumptions.

References

[Airoldi *et al.*, 2008] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

[Aldous, 1985] David Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII*, pages 1–198. 1985.

[Araujo *et al.*, 2014] Miguel Araujo, Stephan Günnemann, Gonzalo Mateos, and Christos Faloutsos. Beyond blocks: Hyperbolic community detection. In *Proc. ECML-PKDD*, pages 50–65, 2014.

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

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proc. ICML*, pages 233–240, 2006.

[Escobar and West, 1994] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90:577–588, 1994.

[Fu *et al.*, 2009] Wenjie Fu, Le Song, and Eric P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proc. ICML*, pages 329–336, 2009.

[Geyer, 2007] Charles J. Geyer. Lower-truncated Poisson and negative binomial distributions. Technical report, Working Paper Written for the Software R. University of Minnesota, MN (available: http://cran.r-project.org/web/packages/aster/vignettes/trunc.pdf), 2007.

[Ho *et al.*, 2011] Qirong Ho, Ankur P. Parikh, Le Song, and Eric P. Xing. Multiscale community blockmodel for network exploration. *Proc. AISTATS*, pages 333–341, 2011.

[Ishiguro *et al.*, 2010] Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua B. Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. In *Proc. NIPS*, pages 919–927, 2010.

[Ishiguro *et al.*, 2012] Katsuhiko Ishiguro, Naonori Ueda, and Hiroshi Sawada. Subset infinite relational models. In *Proc. AISTATS*, pages 547–555, 2012.

[Ishiguro *et al.*, 2016] Katsuhiko Ishiguro, Issei Sato, Masahiro Nakano, Akisato Kimura, and Naonori Ueda. Infinite plaid models for infinite bi-clustering. In *Proc. AAAI*, pages 1701–1708, 2016.

[Kemp *et al.*, 2006] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Proc. AAAI*, volume 1, pages 381–388, 2006.

[Klimat and Yang, 2004] Bryan Klimat and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proc. ECML*, pages 217–226, 2004.

[Liu, 1994] Jun S. Liu. The collapsed Gibbs sampler in bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, 89(427):958–966, 1994.

[Mørup *et al.*, 2011] Morten Mørup, Mikkel N. Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. In *Proc. MLSP*, pages 1–6, 2011.

[MOV, as of 2003] *MovieLens dataset, http://www.grouplens.org/*, as of 2003.

[Nakano *et al.*, 2014] Masahiro Nakano, Katsuhiko Ishiguro, Akisato Kimura, Takeshi Yamada, and Naonori Ueda. Rectangular tiling process. In *Proc. ICML*, pages 361–369, 2014.

[Newman *et al.*, 2009] David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *J. Mach. Learn. Res.*, 10:1801–1828, 2009.

[Nowicki and Snijders, 2001] Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, 96(455):1077–1087, 2001.

[Ohama *et al.*, 2013] Iku Ohama, Hiromi Iida, Takuya Kida, and Hiroki Arimura. An extension of the infinite relational model incorporating interaction between objects. In *Proc. PAKDD (2)*, pages 147–159, 2013.

[Ohama *et al.*, 2015] Iku Ohama, Takuya Kida, and Hiroki Arimura. Multi-layered framework for modeling relationships between biased objects. In *Proc. SDM*, pages 819–827, 2015.

[Ohama *et al.*, 2016] Iku Ohama, Hiromi Iida, Takuya Kida, and Hiroki Arimura. The relevance dependent infinite relational model for discovering co-cluster structure from relationships with structured noise. *IEICE Trans. on Inf. Sys.*, E99-D(4):1139–1152, 2016.

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

[Palla *et al.*, 2012] Konstantina Palla, David A. Knowles, and Zoubin Ghahramani. An infinite latent attribute model for network data. In *Proc. ICML*, pages 1607–1614, 2012.

[Roy *et al.*, 2006] Daniel M. Roy, Charles Kemp, Vikash K. Mansinghka, and Joshua B. Tenenbaum. Learning annotated hierarchies from relational data. In *Proc. NIPS*, pages 1185–1192, 2006.

[Teh *et al.*, 2006a] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.*, 101(476):1566–1581, 2006.

[Teh *et al.*, 2006b] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Proc. NIPS*, pages 1353–1360, 2006.

[Xu *et al.*, 2006] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Infinite hidden relational models. In *Proc. UAI*, pages 544–551, 2006.

[Zhou, 2015] Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Proc. AISTATS*, volume 38, pages 1135–1143, 2015.