Generalized Relational Topic Models with Data Augmentation

Ning Chen[†] **Jun Zhu**[†] **Fei Xia**[‡] **Bo Zhang**[†]
[†]Dept. of CS & T, TNList Lab, State Key Lab of ITS., [‡] School of Software
Tsinghua University, Beijing 100084, China
{ningchen@mail, dcszj@mail, xia-f09@mails, dcszb@mail}.tsinghua.edu.cn

Abstract

Relational topic models have shown promise on analyzing document network structures and discovering latent topic representations. This paper presents three extensions: 1) unlike the common link likelihood with a diagonal weight matrix that allows the-same-topic interactions only, we generalize it to use a full weight matrix that captures all pairwise topic interactions and is applicable to asymmetric networks; 2) instead of doing standard Bayesian inference, we perform regularized Bayesian inference with a regularization parameter to deal with the imbalanced link structure issue in common real networks; and 3) instead of doing variational approximation with strict mean-field assumptions, we present a collapsed Gibbs sampling algorithm for the generalized relational topic models without making restricting assumptions. Experimental results demonstrate the significance of these extensions on improving the prediction performance, and the time efficiency can be dramatically improved with a simple fast approximation method.

1 Introduction

As the availability and scope of network data increase, statistical analysis of such data has attracted a considerable amount of attention. Network data is typically represented as a graph in which the vertices represent entities and edges represent links between entities. Analyzing network data (e.g., link prediction [Liben-Nowell and Kleinberg, 2003; Backstrom and Leskovec, 2011]) could provide useful predictive models for suggesting friends to social network users or citations to scientific articles.

Recent research has focused on latent variable models for link structures, including both parametric [Hoff *et al.*, 2002; Hoff, 2007; Airoldi *et al.*, 2008] and nonparametric Bayesian methods [Miller *et al.*, 2009; Zhu, 2012]. Though modeling network structures well, these models do not account for observed attributes of the entities, such as the text contents of papers in a citation network or the contents of web pages in a hyperlinked network. One work that accounts for both text contents and network structures is relational topic models (RTMs) [Chang and Blei, 2009], an extension of latent Dirichlet allocation (LDA) [Blei *et al.*, 2003] to predict link structures among documents as well as discover their latent topic representations.

Though powerful, existing RTMs have some assumptions that could limit their applicability. First, RTMs define a symmetric link prediction model with a diagonal weight matrix that allows the-same-topic interactions only, and the symmetric nature could make RTMs unsuitable for asymmetric networks. Second, being standard Bayesian models, RTMs do not explicitly deal with the common imbalance issue in real networks which normally have only a few observed links while most entity pairs do not have links. Finally, RTMs and other variants [Liu *et al.*, 2009] apply variational methods to estimate model parameters with normally very strict mean-field assumptions [Jordan *et al.*, 1999].

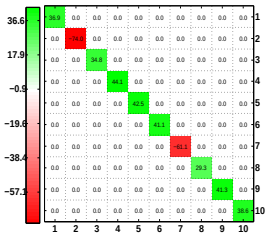
This paper presents three extensions to improve relational topic models: 1) we relax the symmetric assumption and define generalized relational topic models (gRTMs) with a full weight matrix that allows all pairwise topic interactions and is suitable for asymmetric networks; 2) we perform regularized Bayesian inference [Zhu *et al.*, 2011; 2013a] that introduces a regularization parameter to deal with the imbalance problem in common real networks; and 3) we present a collapsed Gibbs sampling algorithm for gRTMs by exploring the classical ideas of data augmentation [Dempster *et al.*, 1977; Tanner and Wong, 1987; Dyk and Meng, 2001]. Technically, we introduce a set of Polya-Gamma random variables [Polson *et al.*, 2012], one per training link, to derive an exact mixture representation of the logistic link likelihood. Then, we can derive the local conditional distributions for collapsed Gibbs sampling analytically. This "augment-and-collapse" algorithm is simple and efficient. More importantly, it does not make restricting assumptions on the desired posterior distribution. Experiments show that these extensions are important and can significantly improve the performance.

The rest paper is structured as follows. Section 2 presents the generalized RTMs. Section 3 presents the "augment-and-collapse" Gibbs sampling algorithm. Section 4 presents experimental results and Section 5 concludes.

2 Generalized Relational Topic Models

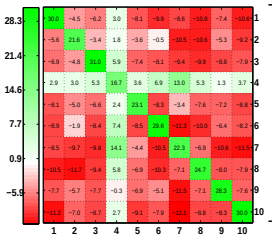
We consider document networks with binary link structures. Let $\mathcal{D} = \{(\mathbf{w}_i, \mathbf{w}_j, y_{ij})\}_{(i,j)\in\mathcal{I}}$ be a labeled training set, where the response variable Y takes values from the output

Table 1: Learned weight matrix of RTM and topics.



learning, bound, PAC, hypothesis, algorithm
numerical, solutions, extensions, approach, remark
mixtures, experts, EM, Bayesian, probabilistic
features, selection, case-based, networks, model
planning, learning, acting, reinforcement, dynamic
genetic, algorithm, evolving, evolutionary, learning
plateau, feature, performance, sparse, networks
modulo, schedule, parallelism, control, processor
neural, cortical, networks, learning, feedforward
markov, models, monte, carlo, Gibbs, sampler

Table 2: Learned weight matrix of gRTM and topics.



genetic, evolving, algorithm, coding, programming
logic, grammars, FOIL, EBG, knowledge, clauses
reinforcement, learning, planning, act, exploration
mixtures, EM, bayesian, networks, learning, genetic
images, visual, scenes, mixtures, networks, learning
decision-tree, rules, induction, learning, features
wake-sleep, learning, networks, cortical, inhibition
monte, carlo, hastings, markov, chain, sampler
case-based, reasoning, CBR, event-based, cases
markov, learning, bayesian, networks, distributions

space $\mathcal{Y} = \{0, 1\}$ and $\mathbf{w}_i = \{w_{in}\}_{n=1}^{N_i}$ denote the words within document i. A relational topic model (RTM) consists of two parts — an LDA model [Blei *et al.*, 2003] for describing the words $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^{D}$ and a classifier for considering link structures $\mathbf{y} = \{y_{ij}\}_{(i,j)\in\mathcal{I}}$. Let K be the number of topics and each topic $\mathbf{\Phi}_k$ is a multinomial distribution over a V-word vocabulary. For Bayesian RTMs, the topics are samples drawn from a prior, e.g., $\mathbf{\Phi}_k \sim \text{Dir}(\boldsymbol{\beta})$, a Dirichlet distribution. The generating process can be described as

1. For each document i
   (a) draw a topic mixing proportion $\boldsymbol{\theta}_i \sim \text{Dir}(\boldsymbol{\alpha})$
   (b) for each word $n = 1, 2, \ldots, N_i$:
      i. draw a topic assignment $z_{in} \sim \text{Mult}(\boldsymbol{\theta}_i)$
      ii. draw the observed word $w_{in} \sim \text{Mult}(\mathbf{\Phi}_{z_{in}})$
2. For each pair of documents i, j:
   (a) draw a link indicator $y_{ij} \sim p(.|\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\eta})$, where $\mathbf{z}_i = \{z_{in}\}_{n=1}^{N_i}$.

$\text{Mult}(\cdot)$ denotes a multinomial distribution; and $\mathbf{\Phi}_{z_{in}}$ denotes the topic selected by the non-zero entry of z_{in}, a K-binary vector with only one entry equaling to 1.

Previous work has defined the link likelihood as

$$p(y_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\eta}) = \sigma(\boldsymbol{\eta}^\top(\bar{\mathbf{z}}_i \circ \bar{\mathbf{z}}_j)), \quad (1)$$

where $\bar{\mathbf{z}}_i = \frac{1}{N_i}\sum_{n=1}^{N_i} \mathbf{z}_{in}$ is the average topic assignments of document i; σ is the sigmoid function; and $\circ$ denotes elementwise product. In [Chang and Blei, 2009], other choices of σ such as the exponential function and the cumulative distribution function of the normal distribution were also used. Here, we focus on the commonly used logistic likelihood model [Miller *et al.*, 2009; Liu *et al.*, 2009].

2.1 The Full RTM Model

Since $\boldsymbol{\eta}^\top(\bar{\mathbf{z}}_i \circ \bar{\mathbf{z}}_j) = \bar{\mathbf{z}}_i^\top \text{diag}(\boldsymbol{\eta})\bar{\mathbf{z}}_j$, the standard RTM learns a diagonal weight matrix which only captures the-same-topic interactions (i.e., there is a non-zero contribution to the link likelihood only when documents i and j have the same topic). One example of the fitted diagonal matrix on the Cora citation network [Chang and Blei, 2009] is shown in Table 1, where each row corresponds to a topic and we show the representative words for the topic at the right hand side. Due to the positiveness restriction of the latent features $\bar{\mathbf{z}}_i$ and the competition between the diagonal entries, some of η_k will have positive values while some are negative. The negative interactions (corresponding to rare topics) may conflict our intuitions of understanding a citation network, where we would expect that papers with the same topics tend to have citation links. Furthermore, by using a diagonal weight matrix, the model is symmetric, i.e., the probability of a link from document i to j is the same as the probability of a link from j to i. The symmetry property does not hold for many networks, e.g., citation networks. To make RTMs more expressive and applicable to asymmetric networks, the first simple extension is to define the link likelihood as

$$p(y_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j, U) = \sigma(\bar{\mathbf{z}}_i^\top U \bar{\mathbf{z}}_j), \quad (2)$$

using a full $K \times K$ weight matrix U. Using the algorithm to be presented, an example of the learned U matrix on the Cora citation network is shown in Table 2. We can see that by allowing all pairwise topic interactions, all the diagonal entries are positive, while most off-diagonal entries are negative. This is consistent with our intuition that papers with the same topics tend to have citation links, while papers with different topics are less likely to have citation links; and there are some papers with generic topics (e.g., topic 4) that have positive link interactions with almost all others.

2.2 Regularized Bayesian Inference

Let $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{D}$ and $\mathbf{\Theta} = \{\boldsymbol{\theta}_i\}_{i=1}^{D}$ denote all the topic assignments and mixing proportions respectively. To fit RTMs, maximum likelihood estimation (MLE) was used with an EM algorithm [Chang and Blei, 2009]. We consider Bayesian inference to get the posterior $p(\mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi}, U|\mathcal{D}) \propto p_0(\mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi}, U)p(\mathcal{D}|\mathbf{Z}, \mathbf{\Phi}, U)$, where $p_0(\mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi}, U) = p_0(U)\big(\prod_i p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\prod_n p(z_{in}|\boldsymbol{\theta}_i)\big)\prod_k p(\mathbf{\Phi}_k|\boldsymbol{\beta})$ is the prior distribution defined by the model and $p(\mathcal{D}|\mathbf{Z}, \mathbf{\Phi}, U) = p(\mathbf{W}|\mathbf{Z}, \mathbf{\Phi})p(\mathbf{y}|\mathbf{Z}, U)$ is the likelihood. One common issue is that real networks are highly imbalanced—the number of positive links is much smaller than the number of negative links. For example, less than 0.1% document pairs in the Cora network have positive links.

To deal with this imbalance issue, we propose to do regularized Bayesian inference [Zhu *et al.*, 2011; 2013a] which offers an extra freedom to handle the imbalance issue in a cost-sensitive manner. Specifically, we define a Gibbs classifier for binary links as follows. If the weight matrix U and topic assignments $\mathbf{Z}$ are given, we build a classifier using the likelihood (2) and the *latent* prediction rule is

$$\hat{y}_{ij}|_{\mathbf{z}_i, \mathbf{z}_j, U} = \mathbb{I}(\sigma(\bar{\mathbf{z}}_i^\top U \bar{\mathbf{z}}_j) > 0.5) = \mathbb{I}(\bar{\mathbf{z}}_i^\top U \bar{\mathbf{z}}_j > 0), \quad (3)$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals to 1 if predicate holds otherwise 0. Since both U and $\mathbf{Z}$ are hidden variables, we infer a posterior distribution $q(U, \mathbf{Z})$ that has the minimal expected log-logistic loss

$$\mathcal{R}(q(U, \mathbf{Z})) = -\sum_{(i,j)\in\mathcal{I}} \mathbb{E}_q[\log p(y_{ij}|\mathbf{z}_i, \mathbf{z}_j, U)], \quad (4)$$

which is a good surrogate loss for the expected link prediction error, $\sum_{ij} \mathbb{E}_q[\mathbb{I}(\hat{y}_{ij}|_{U,\mathbf{z}_i,\mathbf{z}_j} \neq y_{ij})]$, of a Gibbs classifier that randomly draws a model U from the posterior distribution and makes predictions [McAllester, 2003; Germain *et al.*, 2009]. In fact, this choice is supported by the observations that logistic loss has been widely used as a convex surrogate loss for the misclassification error [Rosasco *et al.*, 2004] in the task of binary classification.

With the above Gibbs classifier, we define gRTM as solving the regularized Bayesian inference problem

$$\min_{q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\in\mathcal{P}} \mathcal{L}(q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) + c\mathcal{R}(q(U,\mathbf{Z})) \quad (5)$$

where $\mathcal{L}(q) = \mathrm{KL}(q||p_0(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})]$ is an information theoretical objective; c is a positive regularization parameter controlling the influence from link structures; and $\mathcal{P}$ is the space of normalized distributions. To better understand the above formulation, we define the pseudo-likelihood for links as

$$\psi(y_{ij}|\mathbf{z}_i,\mathbf{z}_j,U) = p^c(y_{ij}|\mathbf{z}_i,\mathbf{z}_j,U) = \frac{\{e^{\omega_{ij}}\}^{cy_{ij}}}{(1+e^{\omega_{ij}})^c}, \quad (6)$$

where $\omega_{ij} = \bar{\mathbf{z}}_i^\top U \bar{\mathbf{z}}_j$ is the discriminant function value. The pseudo-likelihood is un-normalized if $c \neq 1$. Then, problem (5) can be written as

$$\min_{q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\in\mathcal{P}} \mathcal{L}(q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) - \mathbb{E}_q[\log \psi(\mathbf{y}|\mathbf{Z},U)] \quad (7)$$

where $\psi(\mathbf{y}|\mathbf{Z},U) = \prod_{ij}\psi(y_{ij}|\mathbf{z}_i,\mathbf{z}_j,U)$. It can be shown that the optimum solution of problem (5) or the equivalent (7) is the posterior distribution with link information

$$q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = \frac{p_0(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})\psi(\mathbf{y}|\mathbf{Z},U)}{\phi(\mathbf{y},\mathbf{W})}.$$

where $\phi(\mathbf{y},\mathbf{W})$ is the normalization constant to make q as a normalized distribution.

Therefore, by solving problem (5) or (7) we are in fact doing Bayesian inference with a generalized pseudo-likelihood, which is a powered version of the likelihood (2). The flexibility of using regularization parameters can play a significant role in dealing with imbalanced network data as we shall see. For example, we can use a larger c value for the sparse positive links, while using a smaller c for the dense negative links. This simple strategy has been shown effective in learning classifiers [Akbani *et al.*, 2004] and link prediction models [Zhu, 2012] with highly imbalanced data. Finally, an ad hoc generative story can be described as in RTMs, where c can be understood as the pseudo-count of a link.

3 Augment and Collapse Sampling

For the generalized gRTMs or standard RTMs, posterior inference is intractable. Previous solutions use variational techniques with mean-field assumptions. For example, a variational EM algorithm was developed in [Chang and Blei, 2009] with the factorization assumption that $q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = q(U)(\prod_i q(\boldsymbol{\theta}_i)\prod_n q(z_{in}))\prod_k q(\boldsymbol{\Phi}_k)$ which can be too restricted in practice. In this section, we present a simple and efficient Gibbs sampling algorithm without making any restricting assumptions on q. Our "augment-and-collapse" sampling algorithm relies on a data augmentation reformulation of the Bayesian inference problem (7).

3.1 Formulation with Data Augmentation

For the pseudo-likelihood $\psi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta})$, it is not easy to derive a sampling algorithm directly. Instead, we develop our algorithms by introducing auxiliary variables, which lead to a scale mixture of Gaussian components and analytic conditional distributions for Bayesian inference without an accept/reject ratio. Our algorithm represents an extension of Polson's approach [Polson *et al.*, 2012] to deal with non-trivial Bayesian latent variable models for relational data analysis. Let us first introduce the Polya-Gamma variables.

Definition 1 *[Polson* et al.*, 2012] A random variable X has a Polya-Gamma distribution, denoted by $X \sim \mathcal{PG}(a,b)$, if*

$$X = \frac{1}{2\pi^2}\sum_{i=1}^{\infty}\frac{g_i}{(i-1/2)^2 + b^2/(4\pi^2)},$$

where (a,b) are positive parameters and each $g_i \sim \mathcal{G}(a,1)$ is an independent Gamma random variable.

Then, using the ideas of data augmentation [Polson *et al.*, 2012], we have the following results

Lemma 2 *The pseudo-likelihood can be expressed as*

$$\psi(y_{ij}|\mathbf{z}_i,\mathbf{z}_j,U) = \frac{1}{2^c}e^{(\kappa_{ij}\omega_{ij})}\int_0^\infty e^{(-\frac{\lambda_{ij}\omega_{ij}^2}{2})}p(\lambda_{ij}|c,0)d\lambda_d,$$

where $\kappa_{ij} = c(y_{ij}-1/2)$ and λ_{ij} is a Polya-Gamma variable with parameters $a = c$ and $b = 0$.

Lemma 2 indicates that the posterior distribution of gRTMs, i.e., $q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$, can be expressed as the marginal of a higher dimensional distribution that includes the augmented variables $\boldsymbol{\lambda}$. The complete posterior distribution is

$$q(U,\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = \frac{p_0(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})\psi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},U)}{\phi(\mathbf{y},\mathbf{W})},$$

where $\psi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},U) = \prod_{ij}\exp\left(\kappa_{ij}\omega_{ij} - \frac{\lambda_{ij}\omega_{ij}^2}{2}\right)p(\lambda_{ij}|c,0)$ is the pseudo-joint distribution of $\mathbf{y}$ and $\boldsymbol{\lambda}$.

3.2 Inference with Collapsed Gibbs Sampling

Although we can do Gibbs sampling to infer the complete posterior $q(U,\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$ and thus $q(U,\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$ by ignoring $\boldsymbol{\lambda}$, the mixing rate would be slow due to the large sample space. An effective way to reduce the sample space and improve mixing rates is to integrate out the intermediate Dirichlet variables $(\boldsymbol{\Theta},\boldsymbol{\Phi})$ and build a Markov chain whose equilibrium distribution is the collapsed distribution $q(U,\boldsymbol{\lambda},\mathbf{Z})$. Such a collapsed Gibbs sampling procedure has been successfully used in LDA [Griffiths and Steyvers, 2004]. For gRTMs, the collapsed posterior distribution is

$$\begin{aligned} q(U,\boldsymbol{\lambda},\mathbf{Z}) &\propto p_0(U)p(\mathbf{W},\mathbf{Z}|\boldsymbol{\alpha},\boldsymbol{\beta})\psi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},U) \\ &= p_0(U)\prod_{k=1}^{K}\frac{\delta(\mathbf{C}_k+\boldsymbol{\beta})}{\delta(\boldsymbol{\beta})}\prod_{i=1}^{D}\frac{\delta(\mathbf{C}_i+\boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \\ &\quad\times\prod_{ij}\exp\left(\kappa_{ij}\omega_{ij} - \frac{\lambda_{ij}\omega_{ij}^2}{2}\right)p(\lambda_{ij}|c,0), \end{aligned}$$

where $\delta(\mathbf{x}) = \frac{\prod_{i=1}^{\dim(\mathbf{x})}\Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\mathbf{x})}x_i)}$, C_k^t is the number of times the term t being assigned to topic k over the whole corpus and $\mathbf{C}_k = \{C_k^t\}_{t=1}^V$; C_i^k is the number of times that terms being associated with topic k within the i-th document and

$\mathbf{C}_i = \{C_i^k\}_{k=1}^K$. Then, the conditional distributions used in collapsed Gibbs sampling are as follows.

For U: Let $\bar{\mathbf{z}}_{ij} = \text{vec}(\bar{\mathbf{z}}_i \bar{\mathbf{z}}_j^\top)$ and $\boldsymbol{\eta} = \text{vec}(U)$, where $\text{vec}(A)$ is a vector concatenating the row vectors of A. Then we have $\omega_{ij} = \boldsymbol{\eta}^\top \bar{\mathbf{z}}_{ij}$. For the common isotropic Gaussian prior $p_0(U) = \prod_{kk'} \mathcal{N}(U_{kk'}; 0, \nu^2)$, we have

$$q(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}) \propto p_0(\boldsymbol{\eta}) \prod_{ij} \exp\Big(\kappa_{ij} \boldsymbol{\eta}^\top \bar{\mathbf{z}}_{ij} - \frac{\lambda_{ij}(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_{ij})^2}{2}\Big)$$
$$= \mathcal{N}(\boldsymbol{\eta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (8)$$

where the posterior mean is $\boldsymbol{\mu} = \boldsymbol{\Sigma}(\sum_{ij} \kappa_{ij} \bar{\mathbf{z}}_{ij})$ and the covariance is $\boldsymbol{\Sigma} = (\frac{1}{\nu^2} I + \sum_{ij} \lambda_{ij} \bar{\mathbf{z}}_{ij} \bar{\mathbf{z}}_{ij}^\top)^{-1}$. We can easily draw a sample from this K^2-dimensional Gaussian distribution. The inverse can be robustly done using Cholesky decomposition. Since K is normally not large, the inversion is relatively efficient, especially when the number of documents is large.

For Z: The conditional distribution of $\mathbf{Z}$ is

$$q(\mathbf{Z}|U, \boldsymbol{\lambda}) \propto \prod_{k=1}^K \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \prod_{i=1}^D \frac{\delta(\mathbf{C}_i + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \prod_{ij} \psi(y_{ij}|\boldsymbol{\lambda}, \mathbf{Z})$$

where $\psi(y_{ij}|\boldsymbol{\lambda}, \mathbf{Z}) = \exp(\kappa_{ij}\omega_{ij} - \frac{\lambda_{ij}\omega_{ij}^2}{2})$. By canceling common factors, we can derive the local conditional of one variable z_{in} given others $\mathbf{Z}_\neg$ as:

$$q(z_{in}^k = 1|\mathbf{Z}_\neg, U, \boldsymbol{\lambda}, w_{in} = t) \propto \frac{(C_{k,\neg n}^t + \beta_t)(C_{i,\neg n}^k + \alpha_k)}{\sum_t C_{k,\neg n}^t + \sum_{t=1}^V \beta_t}$$
$$\prod_{j \in \mathcal{N}_i^+} \psi(y_{ij}|\boldsymbol{\lambda}, \mathbf{Z}_\neg, z_{in}^k = 1) \prod_{j \in \mathcal{N}_i^-} \psi(y_{ji}|\boldsymbol{\lambda}, \mathbf{Z}_\neg, z_{in}^k = 1), \quad (9)$$

where $C_{\cdot,\neg n}^{\cdot}$ indicates that term n is excluded from the corresponding document or topic; and $\mathcal{N}_i^+ = \{j : (i,j) \in \mathcal{I}\}$ and $\mathcal{N}_i^- = \{j : (j,i) \in \mathcal{I}\}$ denote the neighbors of document i in the training network. We can see that the first term is from the LDA model for observed word counts and the second term is from the link structures $\mathbf{y}$.

For $\boldsymbol{\lambda}$: Finally, the conditional distribution of the augmented variables $\boldsymbol{\lambda}$ is a Polya-Gamma distribution

$$q(\lambda_{ij}|\mathbf{Z}, U) \propto \exp\big(-\frac{\lambda_{ij}\omega_{ij}^2}{2}\big) p(\lambda_{ij}|c, 0) = \mathcal{PG}\big(\lambda_{ij}; c, \omega_{ij}\big). \quad (10)$$

The result is achieved by using the construction definition of the general $\mathcal{PG}(a, b)$ class through an exponential tilting of the $\mathcal{PG}(a, 0)$ density [Polson *et al.*, 2012]. To draw samples from the Polya-Gamma distribution, a naive implementation of the sampling using the infinite sum-of-Gamma representation is not efficient and it also involves a potentially inaccurate step of truncating the infinite sum. Here we adopt the efficient method proposed in [Polson *et al.*, 2012], which draws the samples through drawing samples from the closely related exponentially tilted Jacobi distribution.

With the above conditional distributions, we can construct a Markov chain which iteratively draws samples of $\boldsymbol{\eta}$ (i.e, U) using Eq. (8), $\mathbf{Z}$ using Eq. (9) and $\boldsymbol{\lambda}$ using Eq. (10), with an initial condition. In our experiments, we initially set $\boldsymbol{\lambda} = 1$ and randomly draw $\mathbf{Z}$ from a uniform distribution. In training, we run the Markov chain for M iterations (i.e., the so called burn-in stage). Then, we draw a sample $\hat{U}$ as the final classifier to make predictions on testing data. As we shall see in practice, the Markov chain converges to stable prediction performance with a few burn-in iterations.

3.3 Prediction

Since gRTMs account for both text contents and network structures, we can make predictions for each of them conditioned on the other [Chang and Blei, 2009]. For link prediction, given a test document $\mathbf{w}$, we infer its topic assignments $\mathbf{z}$ in order to apply the classifier (3). This can be done with a collapsed Gibbs sampler, where the conditional distribution is $p(z_n^k = 1|\mathbf{z}_{\neg n}) \propto \hat{\phi}_{kw_n}(C_{\neg n}^k + \alpha_k)$; $C_{\neg n}^k$ is the times that the terms in this document $\mathbf{w}$ assigned to topic k with the n-th term excluded; and $\hat{\boldsymbol{\Phi}}$ is a MAP estimate of the topics, with $\hat{\phi}_{kt} \propto C_k^t + \beta_t$. For word prediction, we infer the distribution $p(w_n|\mathbf{y}, \mathcal{D}, \hat{\boldsymbol{\Phi}}, \hat{U}) = \sum_k \hat{\phi}_{kw_n} p(z_n^k = 1|\mathbf{y}, \mathcal{D}, \hat{U})$. This can be done by drawing a few samples of z_n.

4 Experiments

We present experiments on two public data sets of document networks[1]. The *Cora* data [McCallum *et al.*, 2000] consists of abstracts of 2,708 computer science research papers, with links between documents that cite each other. In total, the Cora citation network has 5,429 positive links, and the dictionary consists of 1,433 words. The *WebKB* data [Craven *et al.*, 1998] contains 877 webpages from the computer science departments of different universities, with links between webpages that are hyper-linked. In total, the WebKB network has 1,608 positive links and the dictionary has 1,703 words.

Since many baseline methods have been outperformed by the standard RTMs in [Chang and Blei, 2009] on the same datasets, we focus on evaluating the effects of the various extensions in the generalized relational topic models (denoted by Gibbs-gRTM) by comparing with its several special cases:

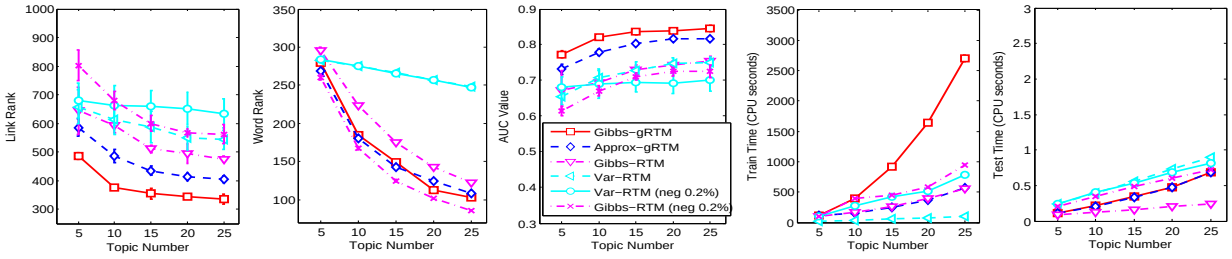
1. **Var-RTM**: the standard RTMs (i.e., $c = 1$) with a diagonal logistic likelihood and a variational EM algorithm with mean-field assumptions [Chang and Blei, 2009];
2. **Gibbs-RTM**: the Gibbs-gRTM model with a diagonal weight matrix for the logistic link likelihood;
3. **Approx-gRTM**: the Gibbs-gRTM model with fast approximation on sampling $\mathbf{Z}$, by computing the link likelihood term in Eq. (8) for once and caching it for sampling all the word topics in each document.

For Var-RTM, we follow the setup [Chang and Blei, 2009] and use positive links only as training data; to deal with the one-class problem, a regularization penalty was used, which in effect injects some number of pseudo-observations (each with a fixed uniform topic distribution). For Gibbs-gRTM models, including Gibbs-RTM and Approx-gRTM, we instead draw some unobserved links as negative examples. Though subsamping normally results in imbalanced datasets, the regularization parameter c in Gibbs-gRTM can effectively address it, as we shall see. Here, we fix c at 1 for negative examples, while we tune it for positive examples.

4.1 Quantitative Results

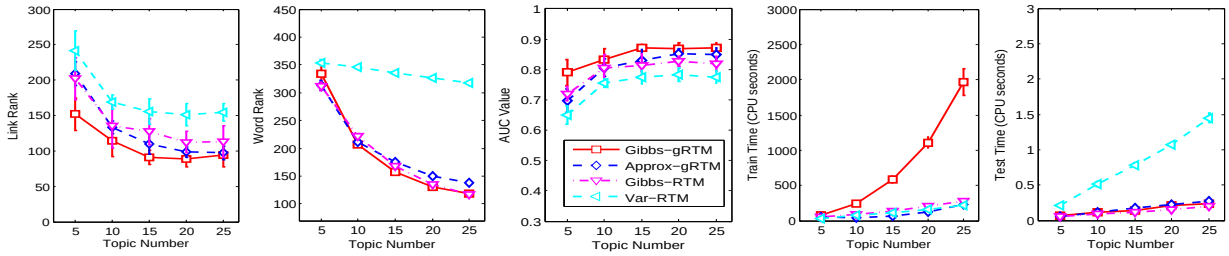
We first report the overall results using the measures of *link rank*, *word rank* and *AUC* (area under ROC curve) of link

[1] http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html



(a) link rank (b) word rank (c) AUC score (d) train time (e) test time

Figure 1: Results of various models with different numbers of topics on the Cora citation data set.



(a) link rank (b) word rank (c) AUC score (d) train time (e) test time

Figure 2: Results of various models with different numbers of topics on the WebKB data set.

prediction, following the setups in [Chang and Blei, 2009]. Link rank is defined as the average rank of the observed links from the held-out test documents to the training documents. The word rank is defined as the average rank of the words in testing documents given their links to the training documents. The test documents are completely new that are not observed during training. In the training phase all the words along with their links of the test documents are removed.

Fig. 1 and Fig. 2 show the 5-fold average results and standard deviations of various models on both datasets with respect to the variation of topic numbers. For the RTM models using collapsed Gibbs sampling, we randomly draw 1% of the unobserved links as negative training examples, which lead to imbalanced training sets. We can see that the generalized Gibbs-gRTM achieves significantly better results on link rank and AUC scores than all other competitors. For word rank, all the RTM models using Gibbs sampling perform better than the RTMs using variational EM methods when the number of topics is larger than 5. The outstanding performance of Gibbs-gRTM is due to many factors. For example, the superior performance of Gibbs-gRTM over the diagonal Gibbs-RTM demonstrates that it is important to consider all pairwise topic interactions to fit real network data; and the superior performance of Gibbs-RTM over Var-RTM shows the benefits of using the regularization parameter c in the regularized Bayesian framework and a collapsed Gibbs sampling algorithm without restricting mean-field assumptions. To single out the influence of the proposed Gibbs sampling algorithm, we also present the results of Var-RTM and Gibbs-RTM with $c = 1$, both of which randomly sample 0.2% unobserved links as negative examples. We can see that by using Gibbs sampling without restricting assumptions, Gibbs-RTM (neg 0.2%) outperforms Var-RTM (neg 0.2%) that makes mean-field assumptions when the number of topics is larger than 10. We defer more careful analysis of other factors in the next section, including c and the subsampling ratio.

Table 3: Split of training time over various steps.

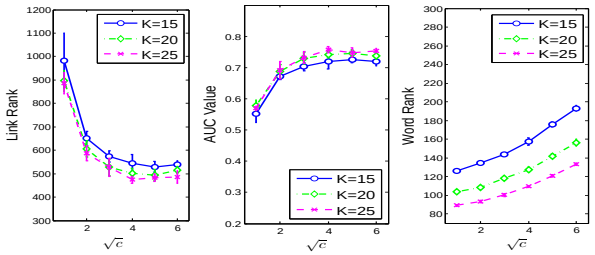
	Sample $\mathbf{Z}$	Sample $\boldsymbol{\lambda}$	Sample U
K=10	331.2 (73.55%)	55.3 (12.29%)	67.8 (14.16%)
K=15	746.8 (76.54%)	55.0 (5.64%)	173.9 (17.82%)
K=20	1300.3 (74.16%)	55.4 (3.16%)	397.7 (22.68%)

We also note that the cost we pay for the outstanding performance of Gibbs-gRTM is on training time, which is much longer than that of Var-RTM because Gibbs-gRTM has K^2 latent features in the logistic likelihood and more training link pairs, while Var-RTM has K latent features and only uses the sparse positive links as training examples. Fortunately, we can apply a simple approximate method in sampling $\mathbf{Z}$ as in Approx-gRTM to significantly improve the training efficiency, while the prediction performance is not sacrificed much. In fact, Approx-gRTM is still significantly better than Var-RTM in all cases, and it has comparable link prediction performance with Gibbs-gRTM on the WebKB dataset, when K is large. Table 3 further shows the training time spent on each sub-step of the Gibbs sampling algorithm of Gibbs-gRTM. We can see that the step of sampling $\mathbf{Z}$ takes most of the time ($> 70\%$); and the steps of sampling $\mathbf{Z}$ and $\boldsymbol{\eta}$ take more time as K increases, while the step of sampling $\boldsymbol{\lambda}$ takes almost a constant time when K changes.

4.2 Sensitivity Analysis

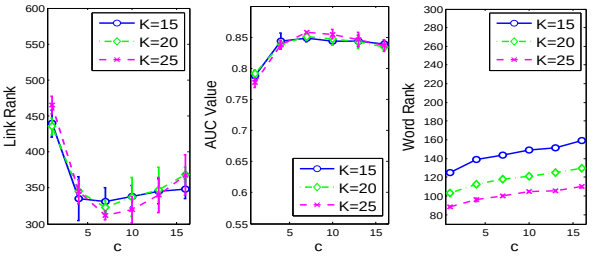
To get insights about the outstanding performance of Gibbs-gRTM, we present a careful analysis of various factors.

Hyper-parameters c: Fig. 3 and Fig. 4 show the prediction performance of the diagonal Gibbs-RTM and the generalized Gibbs-gRTM on the Cora dataset with different c values. For Gibbs-RTM, we can see that the link rank decreases and AUC scores increase when c becomes larger and the prediction performance is stable in a wide range (e.g., $2 \leq \sqrt{c} \leq 6$). But the RTM model (i.e., $c = 1$) using Gibbs sampling doesn't



(a) link rank (b) AUC score (c) word rank

Figure 3: Performance of Gibbs-RTM with different c values on the Cora dataset.



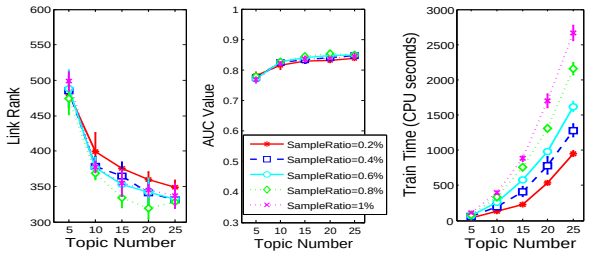
(a) link rank (b) AUC score (c) word rank

Figure 4: Performance of Gibbs-gRTM with different c values on the Cora dataset.

perform well due to its ineffectiveness in dealing with highly imbalanced network data. In Fig. 4, we can observe that when $4 \leq c \leq 10$, the link rank and AUC scores of Gibbs-gRTM achieve the local optimum, which performs much better than the performance of Gibbs-gRTM when $c = 1$. By comparing Fig. 3 and Fig. 4, we can see that Gibbs-gRTM generally needs a smaller c to get the best performance. This is because by allowing all pairwise topic interactions, Gibbs-gRTM is much more expressive than Gibbs-RTM; and thus easier to over-fit when c gets large. For both Gibbs-RTM and Gibbs-gRTM, the word rank increases slowly with the growth of c. This is because a larger c value makes the model more concentrated on fitting link structures and thus the fitness of observed words sacrifices a bit. But if we compare with the variational RTM (i.e., Var-RTM) as shown in Fig. 1, the word ranks of both Gibbs-RTM and Gibbs-gRTM are much lower for all the c values we have tested. This suggests the advantages of the collapsed Gibbs sampling algorithm. In the previous experiments, we has set $c = 25$ for Gibbs-RTM and $c = 4$ for Gibbs-gRTM.

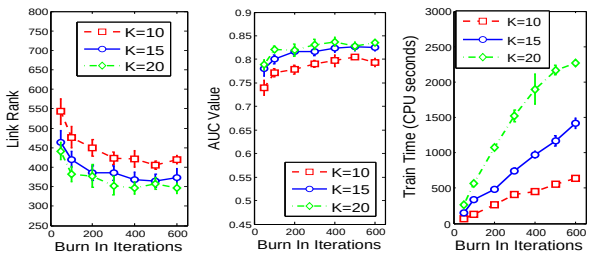
Subsample ratio: We analyze the influence of the subsample ratio on the performance of Gibbs-gRTM on the Cora data in Fig. 5. In total, less than 0.1% links are positive on the Cora networks. We can see that by introducing the regularization parameter c, Gibbs-gRTM can effectively fit various imbalanced network data and the different subsample ratios have a weak influence on the performance of Gibbs-gRTM. Since a larger subsample ratio leads to a bigger training set, the training time increases as expected.

Burn-In: We analyze the sensitivity of Gibbs-gRTM on Cora dataset with respect to the number of burn-in iterations. Fig. 6 show the performance of Gibbs-gRTM with different numbers of burn-in iterations. We can see that the link rank and AUC scores converge fast to stable optimum points with



(a) link rank (b) AUC score (c) train time

Figure 5: Performance of Gibbs-gRTM with different numbers of negative training links on the Cora dataset.



(a) link rank (b) AUC score (c) train time

Figure 6: Performance of Gibbs-gRTM with different burn-in steps for Cora dataset.

about 300 iterations. The training time grows almost linearly with respect to the number of burn-in iterations. We have similar observations for the diagonal Gibbs-RTM model and Approx-gRTM with fast approximation. In all the previous experiments, we have set the burn-in steps at 400.

Finally, Gibbs-gRTM models are insensitive to the other parameters (e.g., the Dirichlet prior α) and we omit the details due to space limitation.

5 Conclusions and Discussions

We have presented a generalized relational topic model for considering all pairwise topic interactions and dealing with imbalanced network data by doing regularized Bayesian inference. We also presented a simple "augment-and-collapse" sampling algorithm without restricting assumptions on the posterior distribution. Experiments on real network data demonstrate significant improvements on prediction tasks. The time efficiency can be significantly improved with a simple approximation method.

For future work, we are interested in making the sampling algorithm scalable to large networks by using distributed architectures [Smola and Narayanamurthy, 2010] or doing online inference [Hoffman *et al.*, 2010]. Moreover, the data augmentation idea can be applied to solve the posterior inference problem of other Bayesian logistic latent variable models, such as supervised topic models [Zhu *et al.*, 2013b].

Acknowledgments

This work is supported by National Key Project for Basic Research of China (Grant Nos: 2013CB329403, 2012CB316301), Tsinghua Self-innovation Project (Grant Nos: 20121088071, 20111081111), and China Postdoctoral Science Foundation Grant (Grant No: 2012M520281).

References

[Airoldi *et al.*, 2008] E. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, 2008.

[Akbani *et al.*, 2004] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *European Conference on Machine Learning*, 2004.

[Backstrom and Leskovec, 2011] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *ACM International Conference on Web Search and Data Mining*, 2011.

[Blei *et al.*, 2003] D. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, 2003.

[Chang and Blei, 2009] J. Chang and D. Blei. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[Craven *et al.*, 1998] M. Craven, D. Dipasquo, D. Freitag, and A. McCallum. Learning to extract symbolic knowledge from the world wide web. In *AAAI Conference on Artificial Intelligence*, 1998.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Ser. B*, (39):1–38, 1977.

[Dyk and Meng, 2001] D. Van Dyk and X. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[Germain *et al.*, 2009] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, pages 353–360, 2009.

[Griffiths and Steyvers, 2004] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.

[Hoff *et al.*, 2002] P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of American Statistical Association*, 97(460), 2002.

[Hoff, 2007] P.D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, 2007.

[Hoffman *et al.*, 2010] M. Hoffman, D. Blei, and F. Bach. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2010.

[Jordan *et al.*, 1999] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. *An introduction to variational methods for graphical models*. MIT Press, Cambridge, MA, 1999.

[Liben-Nowell and Kleinberg, 2003] D. Liben-Nowell and J.M. Kleinberg. The link prediction problem for social networks. In *ACM Conference of Information and Knowledge Management*, 2003.

[Liu *et al.*, 2009] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link LDA: Joint models of topic and author community. In *International Conference on Machine Learning*, 2009.

[McAllester, 2003] D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

[McCallum *et al.*, 2000] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 2000.

[Miller *et al.*, 2009] K. Miller, T. Griffiths, and M. Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, 2009.

[Polson *et al.*, 2012] N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *arXiv:1205.0310v1*, 2012.

[Rosasco *et al.*, 2004] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, (16):1063–1076, 2004.

[Smola and Narayanamurthy, 2010] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. *International Conference on Very Large Data Bases*, 2010.

[Tanner and Wong, 1987] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the Americal Statistical Association*, 82(398):528–540, 1987.

[Zhu *et al.*, 2011] J. Zhu, N. Chen, and E.P. Xing. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems*, 2011.

[Zhu *et al.*, 2013a] J. Zhu, N. Chen, and E.P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. In *arXiv:1210.1766v2*, 2013.

[Zhu *et al.*, 2013b] J. Zhu, X. Zheng, and B. Zhang. Bayesian logistic supervised topic models with data augmentation. In *The Annual Meeting of the Association for Computational Linguistics*, 2013.

[Zhu, 2012] J. Zhu. Max-margin nonparametric latent feature models for link prediction. In *International Conference on Machine Learning*, 2012.