Relational Topic Models for Document Networks

Jonathan Chang
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
jcone@princeton.edu

David M. Blei
Department of Computer Science
Princeton University
35 Olden St.
Princeton, NJ 08544
blei@cs.princeton.edu

Abstract

We develop the relational topic model (RTM), a model of documents and the links between them. For each pair of documents, the RTM models their link as a binary random variable that is conditioned on their contents. The model can be used to summarize a network of documents, predict links between them, and predict words within them. We derive efficient inference and learning algorithms based on variational methods and evaluate the predictive performance of the RTM for large networks of scientific abstracts and web documents.

1 INTRODUCTION

Network data, such as citation networks of documents, hyperlinked networks of web pages, and social networks of friends, are becoming pervasive in modern machine learning applications. Analyzing network data provides useful predictive models, pointing social network members towards new friends, scientific papers towards relevant citations, and web pages towards other related pages.

Recent research in this field has focused on latent variable models of link structure, models which decompose a network according to hidden patterns of connections between its nodes (Kemp et al. 2004; Hoff et al. 2002; Hofman and Wiggins 2007; Airoldi et al. 2008). Though powerful, these models account only for the structure of the network, ignoring observed attributes of the nodes. For example, a network model can find patterns which account for the citation connections between scientific articles, but it cannot also account for the texts.

This type of information about the nodes, along with the links between them, should be used for uncovering, understanding and exploiting the latent structure in the data. To this end, we develop a new model of network data that accounts for both links such as citations and attributes such as text.

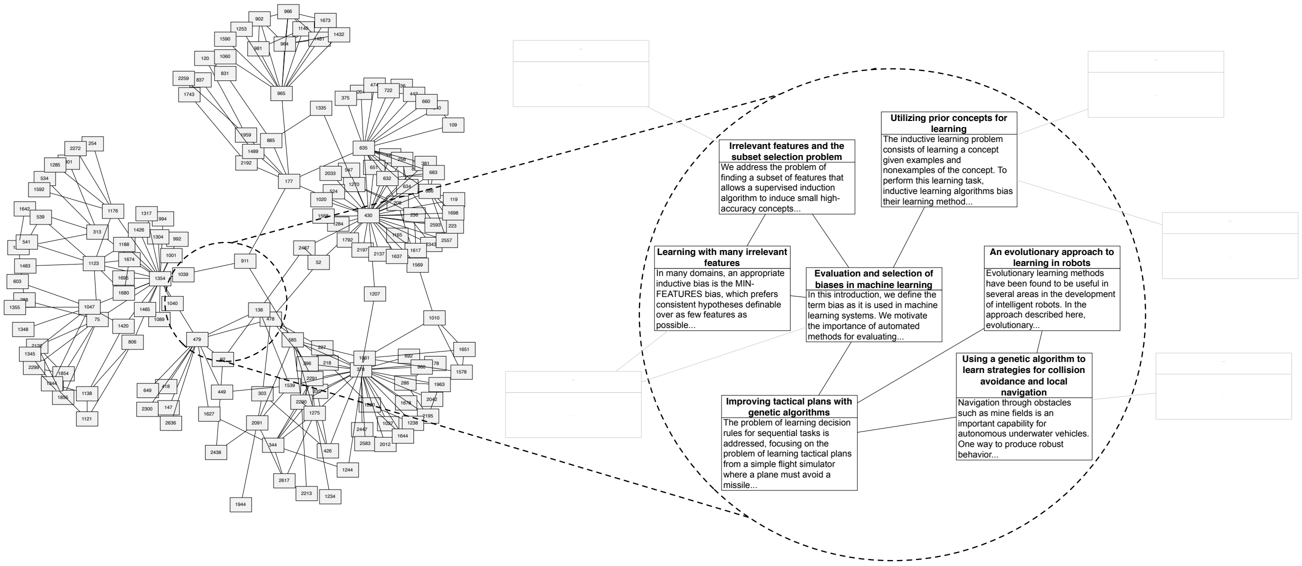
Accounting for patterns in both sources of data leads to a more powerful model than those that only consider links. Given a new node and some of its links, traditional models of network structure can provide a predictive distribution of other nodes with which it it might be connected. Our model need not observe any links of a new node; it can predict links using only its attributes. Thus, we can suggest citations of newly written papers, predict the likely hyperlinks of a web page in development, or suggest friendships in a social network based only on a new user's profile of interests. Moreover, given a new node and its links, our model provides a predictive distribution of node attributes. This complementary predictive mechanism can be used to predict keywords from citations or a user's interests from his or her social connections. These types of predictions are out of reach for traditional network models.

Our model is the *relational topic model* (RTM), a hierarchical model of links and node attributes. Focusing on networks of text data, the RTM explicitly ties the content of the documents with the connections between them. First, we describe the statistical assumptions behind the RTM. Then, we derive efficient algorithms for approximate posterior inference, parameter estimation, and prediction. Finally, we study its performance on scientific citation networks and hyperlinked web pages. The RTM provides significantly better word prediction and link prediction than natural alternatives and the current state of the art.

2 RELATIONAL TOPIC MODELS

The *relational topic model* (RTM) is a model of data composed of documents, which are collections of words, and links between them (see Figure 1). It embeds this data in a latent space that explains both the words of the documents and how they are connected.

Appearing in Proceedings of the 12^{th} International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.



Irrelevant features and the subset selection problem

We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...

Utilizing prior concepts for learning

The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...

Learning with many irrelevant features

In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...

Evaluation and selection of biases in machine learning

In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...

An evolutionary approach to learning in robots

Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...

Using a genetic algorithm to learn strategies for collision avoidance and local navigation

Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...

Improving tactical plans with genetic algorithms

The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

Figure 1: Example data appropriate for the relational topic model. Each document is represented as a bag of words and linked to other documents via citation. The RTM defines a joint distribution over the words in each document and the citation links between them.

The RTM is based on latent Dirichlet allocation (LDA) (Blei et al. 2003). LDA is a generative probabilistic model that uses a set of "topics," distributions over a fixed vocabulary, to describe a corpus of documents. In its generative process, each document is endowed with a Dirichlet-distributed vector of topic proportions, and each word of the document is assumed drawn by first drawing a topic assignment from those proportions and then drawing the word from the corresponding topic distribution.

In the RTM, each document is first generated from topics as in LDA. The links between documents are then modeled as binary variables, one for each pair of documents. These are distributed according to a distribution that depends on the topics used to generate each of the constituent documents. In this way, the content of the documents are statistically connected to the link structure between them.

The parameters to the RTM are K distributions over terms $\beta_{1:K}$, a K-dimensional Dirichlet parameter α, and a function ψ that provides binary probabilities. (This function is explained in detail below.) The RTM assumes that a set of observed documents $w_{1:D,1:N}$ and binary links between them $y_{1:D,1:D}$ are generated by the following process.

1. For each document d:
   (a) Draw topic proportions $\theta_d|\alpha \sim \mathrm{Dir}(\alpha)$.
   (b) For each word $w_{d,n}$:
      i. Draw assignment $z_{d,n}|\theta_d \sim \mathrm{Mult}(\theta_d)$.
      ii. Draw word $w_{d,n}|z_{d,n}, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_{d,n}})$.
2. For each pair of documents d, d':
   (a) Draw binary link indicator

$$y|\boldsymbol{z}_d, \boldsymbol{z}_{d'} \sim \psi(\cdot|\boldsymbol{z}_d, \boldsymbol{z}_{d'}).$$

Figure 2 illustrates the graphical model for this process for a single pair of documents. The full model, which is difficult to illustrate, contains the observed words from all D documents, and D^2 link variables for each possible connection between them.

The function ψ is the *link probability function* that defines a distribution over the link between two documents. This function is dependent on the topic assignments that generated their words, $\boldsymbol{z}_d$ and $\boldsymbol{z}_{d'}$. We explore two possibilities.

First, we consider

$$\psi_\sigma(y=1) = \sigma(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}) + \nu), \tag{1}$$

where $\overline{\mathbf{z}}_d = \frac{1}{N_d}\sum_n z_{d,n}$, the $\circ$ notation denotes the Hadamard (element-wise) product, and the function σ is the sigmoid. This link function models each per-pair binary variable as a logistic regression with hidden covariates. It is parameterized by coefficients η and intercept ν. The covariates are constructed by the Hadamard product of $\overline{\mathbf{z}}_d$ and $\overline{\mathbf{z}}_{d'}$, which captures similarity between the hidden topic representations of the two documents.

Second, we consider

$$\psi_e(y=1) = \exp(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}) + \nu). \tag{2}$$

Here, ψ_e uses the same covariates as ψ_σ, but has an exponential mean function instead. Rather than tapering off when $\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}$ are close, the probabilities returned by this function continue to increases exponentially. With some algebraic manipulation, the function ψ_e can be viewed as an approximate variant of the modeling methodology presented in Blei and Jordan (2003).

In both of the ψ functions we consider, the response is a function of the latent feature expectations, $\overline{\mathbf{z}}_d$ and $\overline{\mathbf{z}}_{d'}$. This

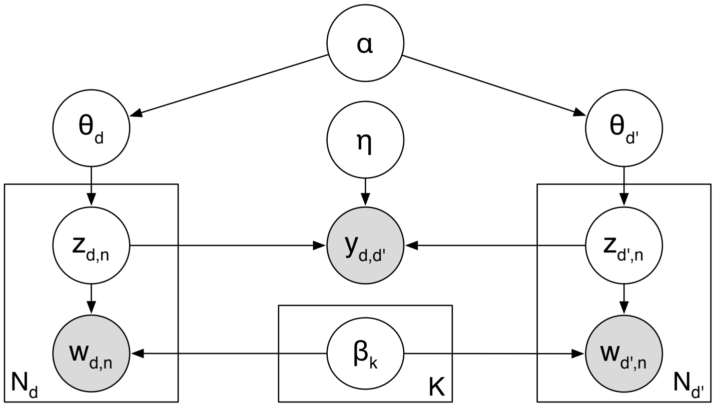


Figure 2: A two-document segment of the RTM. The variable y indicates whether the two documents are linked. The complete model contains this variable for each pair of documents. The plates indicate replication. This model captures both the words and the link structure of the data shown in Figure 1.

formulation, inspired by the supervised LDA model (Blei and McAuliffe 2007), ensures that the same latent topic assignments used to generate the content of the documents also generates their link structure. Models which do not enforce this coupling, such as Nallapati et al. (2008), might divide the topics into two independent subsets—one for links and the other for words. Such a decomposition prevents these models from making meaningful predictions about links given words and words given links. In Section 4 we demonstrate empirically that the RTM outperforms such models on these tasks.

3 INFERENCE, ESTIMATION, AND PREDICTION

With the model defined, we turn to approximate posterior inference, parameter estimation, and prediction. We develop a variational inference procedure for approximating the posterior. We use this procedure in a variational expectation-maximization (EM) algorithm for parameter estimation. Finally, we show how a model whose parameters have been estimated can be used as a predictive model of words and links.

Inference In posterior inference, we seek to compute the posterior distribution of the latent variables conditioned on the observations. Exact posterior inference is intractable (Blei et al. 2003; Blei and McAuliffe 2007). We appeal to variational methods.

In variational methods, we posit a family of distributions over the latent variables indexed by free variational parameters. Those parameters are fit to be close to the true posterior, where closeness is measured by relative entropy. See Jordan et al. (1999) for a review. We use the fully-factorized family,

$$q(\mathbf{\Theta}, \mathbf{Z}|\boldsymbol{\gamma}, \mathbf{\Phi}) = \textstyle\prod_d \left[q_\theta(\theta_d|\gamma_d) \prod_n q_z(z_{d,n}|\phi_{d,n}) \right], \quad (3)$$

where $\boldsymbol{\gamma}$ is a set of Dirichlet parameters, one for each document, and $\mathbf{\Phi}$ is a set of multinomial parameters, one for each word in each document. Note that $\mathbb{E}_q \left[z_{d,n} \right] = \phi_{d,n}$.

Minimizing the relative entropy is equivalent to maximizing the Jensen's lower bound on the marginal probability of the observations, i.e., the evidence lower bound (ELBO),

$$\begin{aligned}\mathscr{L} = {} & \textstyle\sum_{(d_1,d_2)} \mathbb{E}_q \left[\log p(y_{d_1,d_2}|\boldsymbol{z}_{d_1}, \boldsymbol{z}_{d_2}, \boldsymbol{\eta}, \nu) \right] + \\ & \textstyle\sum_d \sum_n \mathbb{E}_q \left[\log p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right] + \\ & \textstyle\sum_d \sum_n \mathbb{E}_q \left[\log p(z_{d,n}|\theta_d) \right] + \\ & \textstyle\sum_d \mathbb{E}_q \left[\log p(\theta_d|\alpha) \right] + \mathrm{H}(q), \end{aligned} \quad (4)$$

where (d_1, d_2) denotes all document pairs. The first term of the ELBO differentiates the RTM from LDA (Blei et al. 2003). The connections between documents affect the objective in approximate posterior inference (and, below, in parameter estimation).

We develop the inference procedure under the assumption that only observed links will be modeled (i.e., y_{d_1,d_2} is either 1 or unobserved).[1] We do this for two reasons.

First, while one can fix $y_{d_1,d_2} = 1$ whenever a link is observed between d_1 and d_2 and set $y_{d_1,d_2} = 0$ otherwise, this approach is inappropriate in corpora where the absence of a link cannot be construed as evidence for $y_{d_1,d_2} = 0$. In these cases, treating these links as unobserved variables is more faithful to the underlying semantics of the data. For example, in large social networks such as Facebook the absence of a link between two people does not necessarily mean that they are not friends; they may be real friends who are unaware of each other's existence in the network. Treating this link as unobserved better respects our lack of knowledge about the status of their relationship.

Second, treating non-links links as hidden decreases the computational cost of inference; since the link variables are leaves in the graphical model they can be removed when-

[1] Sums over document pairs (d_1, d_2) are understood to range over pairs for which a link has been observed.

ever they are unobserved. Thus the complexity of computation scales with the number of observed links rather than the number of document pairs. This provides a significant computational advantage.

Our aim now is to compute each term of the objective function given in Equation 4. The first term depends on our choice of link probability function. This term is not tractable to compute when the logistic function of Equation 1 is chosen. We use a first-order approximation (Braun and McAuliffe 2007),

$$\begin{aligned}\mathscr{L}_{d_1,d_2} \equiv \mathbb{E}_q\left[\log p(y_{d_1,d_2} = 1 | \boldsymbol{z}_{d_1}, \boldsymbol{z}_{d_2}, \boldsymbol{\eta}, \nu)\right] \approx \\ \boldsymbol{\eta}^T \overline{\boldsymbol{\pi}}_{d_1,d_2} + \nu + \log \sigma\left(-\boldsymbol{\eta}^T \overline{\boldsymbol{\pi}}_{d_1,d_2} - \nu\right), \qquad (5)\end{aligned}$$

where $\overline{\boldsymbol{\pi}}_{d_1,d_2} = \overline{\boldsymbol{\phi}}_{d_1} \circ \overline{\boldsymbol{\phi}}_{d_2}$ and $\overline{\boldsymbol{\phi}}_d = \mathbb{E}_q\left[\overline{\mathbf{z}}_d\right] = \frac{1}{N_d}\sum_n \phi_{d,n}$. When ψ_e is the response function, this term can be computed explicitly as

$$\mathbb{E}_q\left[\log p(y_{d_1,d_2} = 1 | \overline{\mathbf{z}}_{d_1}, \overline{\mathbf{z}}_{d_2}, \boldsymbol{\eta}, \nu)\right] = \boldsymbol{\eta}^T \overline{\boldsymbol{\pi}}_{d_1,d_2} + \nu. \quad (6)$$

We use coordinate ascent to optimize the ELBO with respect to the variational parameters $\boldsymbol{\gamma}, \boldsymbol{\Phi}$,

$$\begin{aligned}\phi_{d,j} \propto \exp\{\textstyle\sum_{d' \neq d}(\nabla_{\boldsymbol{\pi}_{d,d'}} \mathscr{L}_{d,d'}) \frac{\boldsymbol{\eta} \circ \overline{\boldsymbol{\phi}}_{d'}}{N_d} + \\ \mathbb{E}_q\left[\log \theta_d | \gamma_d\right] + \log \beta_{\cdot, w_{d,j}}\},\end{aligned}$$

where $\mathscr{L}_{d,d'}$ is computed according to either Equation 5 or 6 depending on the choice of ψ. $\log \beta_{\cdot, w_{d,j}}$ can be computed by taking the element-wise logarithm of the $w_{d,j}$th column of $\boldsymbol{\beta}$. $\mathbb{E}_q\left[\log \boldsymbol{\theta}_d | \boldsymbol{\gamma}_d\right]$ is $\Psi(\boldsymbol{\gamma}_d) - \Psi(\sum \gamma_{d,i})$, where Ψ is the digamma function. (A digamma of a vector is the vector of digammas.)

The update for $\boldsymbol{\gamma}$ is identical to that in variational inference for LDA (Blei et al. 2003), $\gamma_d \leftarrow \alpha + \sum_n \boldsymbol{\phi}_{d,n}$.

Parameter estimation We fit the model by finding maximum likelihood estimates for each of the parameters: multinomial topic vectors $\beta_{1:K}$ and link function parameters $\boldsymbol{\eta}, \nu$. Once again, this is intractable so we turn to an approximation. We employ variational expectation-maximization, where we iterate between optimizing the ELBO of Equation 4 with respect to the variational distribution and with respect to the model parameters.

Optimizing with respect to the variational distribution is described in Section 3. Optimizing with respect to the model parameters is equivalent to maximum likelihood estimation with expected sufficient statistics, where the expectation is taken with respect to the variational distribution.

Since the terms in Equation 4 that involve $\boldsymbol{\beta}$ are identical to those in LDA, estimating the topic vectors can be done via the same update:

$$\beta_{k,w} \propto \sum_d \sum_n \mathbb{1}(w_{d,n} = w)\phi^k_{d,n}.$$

In practice, we smooth our estimates of $\beta_{k,w}$ using a symmetric Dirichlet prior on the topics.

It is not possible to directly optimize the parameters of the link probability function without negative observations (i.e., $y_{d_1,d_2} = 0$). We address this by applying a regularization penalty parameterized by a scalar, ρ. The effect of this regularization is to posit some number of latent negative observations in the network and to incorporate them into the parameter estimates. The frequency of the negative observations is controlled by ρ. (For space we omit the derivation of this regularization term.)

When using the logistic function of Equation 1, we use gradient-based optimization to estimate the parameters $\boldsymbol{\eta}$ and ν. Using the approximation used in Equation 5, the relevant gradients of the ELBO are

$$\begin{aligned}\nabla_{\boldsymbol{\eta}} \mathscr{L} &\approx \textstyle\sum_{(d_1,d_2)}\left[1 - \sigma\left(\boldsymbol{\eta}^{\mathrm{T}} \overline{\boldsymbol{\pi}}_{d_1,d_2} + \nu\right)\right] \overline{\boldsymbol{\pi}}_{d_1,d_2} - \\ &\quad \rho\sigma\left(\boldsymbol{\eta}/K^2 + \nu\right)/K^2, \\ \tfrac{\partial}{\partial \nu} \mathscr{L} &\approx \textstyle\sum_{(d_1,d_2)}\left[1 - \sigma\left(\boldsymbol{\eta}^{\mathrm{T}} \overline{\boldsymbol{\pi}}_{d_1,d_2} + \nu\right)\right] - \\ &\quad \rho\sigma\left(\mathbf{1}^{\mathrm{T}} \eta/K^2 + \nu\right).\end{aligned}$$

When using the exponential function of Equation 2, we can estimate the parameters $\boldsymbol{\eta}$ and ν analytically,

$$\begin{aligned}\nu &\leftarrow \log\left(1 - \mathbf{1}^{\mathrm{T}} \bar{\boldsymbol{\Pi}}\right) - \log\left(\rho \frac{K-1}{K} + 1 - \mathbf{1}^{\mathrm{T}} \bar{\boldsymbol{\Pi}}\right) \\ \eta &\leftarrow \log\left(\bar{\boldsymbol{\Pi}}\right) - \log\left(\bar{\boldsymbol{\Pi}} + \frac{\rho}{K^2}\mathbf{1}\right) - \mathbf{1}\nu,\end{aligned}$$

where $\bar{\boldsymbol{\Pi}} = \sum_{(d_1,d_2)} \overline{\boldsymbol{\pi}}_{d_1,d_2}$.

Prediction With a fitted model, our ultimate goal is to make predictions about new data. We describe two kinds of prediction: link prediction from words and word prediction from links.

In link prediction, we are given a new document (i.e. a document which is not in the training set) and its words. We are asked to predict its links to the other documents. This requires computing

$$\begin{aligned}p(y_{d,d'} | \boldsymbol{w}_{\boldsymbol{d}}, \boldsymbol{w}_{\boldsymbol{d'}}) = \\ \textstyle\sum_{\boldsymbol{z}_d, \boldsymbol{z}_{d'}} p(y_{d,d'} | \overline{\mathbf{z}}_d, \overline{\mathbf{z}}_{d'}) p(\boldsymbol{z}_d, \boldsymbol{z}_{d'} | \boldsymbol{w}_{\boldsymbol{d}}, \boldsymbol{w}_{\boldsymbol{d'}}),\end{aligned}$$

an expectation with respect to a posterior that we cannot compute. Using the inference algorithm from Section 3, we find variational parameters which optimize the ELBO for the given evidence, i.e., the words and links for the training documents and the words in the test document. Replacing the posterior with this approximation $q(\boldsymbol{\Theta}, \boldsymbol{Z})$, the predictive probability is approximated with

$$p(y_{d,d'} | \boldsymbol{w}_{\boldsymbol{d}}, \boldsymbol{w}_{\boldsymbol{d'}}) \approx \mathbb{E}_q\left[p(y_{d,d'} | \overline{\mathbf{z}}_d, \overline{\mathbf{z}}_{d'})\right]. \qquad (7)$$

In a variant of link prediction, we are given a new set of documents (documents not in the training set) along with their words and asked to select the links most likely to exist. The predictive probability for this task is proportional to Equation 7.

The second predictive task is word prediction, where we predict the words of a new document based only on its links. As with link prediction, $p(w_{d,i}|\boldsymbol{y_d})$ cannot be computed. Using the same technique, a variational distribution can approximate this posterior. This yields the predictive probability

$$p(w_{d,i}|\boldsymbol{y_d}) \approx \mathbb{E}_q \left[p(w_{d,i}|z_{d,i}) \right].$$

Note that models which treat the endpoints of links as lexical tokens cannot participate in the two tasks presented here because they cannot make meaningful predictions for documents that do not appear in the training set (Nallapati and Cohen 2008; Cohn and Hofmann 2001; Sinkkonen et al. 2008). By modeling both documents and links generatively, our model is able to give predictive distributions for words given links, links given words, or any mixture thereof.

4 EMPIRICAL RESULTS

We examined the RTM on three data sets. Words were stemmed; stop words and infrequently occurring words were removed. Directed links were converted to undirected links[2] and documents with no links were removed. The *Cora* data (McCallum et al. 2000) contains abstracts from the Cora research paper search engine, with links between documents that cite each other. The *WebKB* data (Craven et al. 1998) contains web pages from the computer science departments of different universities, with links determined from the hyperlinks on each page. The *PNAS* data contains recent abstracts from the Proceedings of the National Academy of Sciences. The links between documents are intra-*PNAS* citations.[3]

Evaluating the predictive distribution As with any probabilistic model, the RTM defines a probability distribution over unseen data. After inferring the latent variables from data (as described in Section 3), we ask how well the model predicts the links and words of unseen nodes. Models that give higher probability to the unseen documents better capture the joint structure of words and links.

We study the two variants of the RTM discussed above: logistic RTM uses the logistic link of Equation 1; exponential RTM uses the exponential link of Equation 2. We compare these models against three alternative approaches. The first ("Baseline") models words and links independently. The words are modeled with a multinomial; the links are modeled with a Bernoulli. The second ("Mixed-Membership") is the model proposed by Nallapati et al. (2008), which is an extension of the mixed membership stochastic block model (Airoldi et al. 2008) to model network structure and node attributes. The third ("LDA + Regression") first fits an LDA model to the documents and then fits a logistic regression model to the observed links, with input given by the Hadamard product of the latent class distributions of each pair of documents. Rather than performing dimensionality reduction and regression simultaneously, this method performs unsupervised dimensionality reduction first, and then regresses to understand the relationship between the latent space and underlying link structure. All models were trained such that the total mass of the Dirichlet hyperparameter α was 5.0. (While we omit a full sensitivity study here, we observed that the performance of the models was similar for α within a factor of 2 above and below the value we chose.)

We measured the performance of these models on link prediction and word prediction (see Section 3). We divided each data set into five folds. For each fold and for each model, we ask two predictive queries: given the words of a new document, what is the likelihood of its links; and given the links of a new document, what is the likelihood of its words? Again, the predictive queries are for completely new test documents that are not observed in training. During training the test documents are removed along with their attendant links. We show the results for both tasks in Figure 3.

In predicting links, the two variants of the RTM perform better than all of the alternative models for all of the data sets (see Figure 3, top row). *Cora* is paradigmatic, showing a nearly 6% improvement in log likelihood for exponential RTM over baseline and 5% improvement over LDA + Regression. Logistic RTM performs nearly as well on *Cora* with an approximately 5% improvement over baseline and 4% improvement over LDA + Regression. We emphasize that the links are predicted to documents seen in the training set from documents which were held out. By incorporating link and node information in a joint fashion, the model is able to generalize to new documents for which no link information was previously known.

The performance of the Mixed-Membership model rarely deviates from the baseline. Despite its increased dimensionality (and commensurate increase in computational difficulty), only on *PNAS* and only when the number of topics is large is the Mixed-Membership model competitive with any of the proposed models. We hypothesize that the Mixed-Membership model exhibits this behavior because it uses some topics to explain the words observed in the train-

[2] The RTM can be extended to accommodate directed connections. Here we modeled undirected links.

[3] After processing, the *Cora* data contained 2708 documents, 49216 words, 5278 links, and a lexicon of 1433 terms. The *WebKB* data contained 877 documents, 79365 words, 1388 links, and a lexicon of 1703 terms. The *PNAS* data contained 2128 documents, 119162 words, 1577 links, and had a lexicon of 2239 terms.

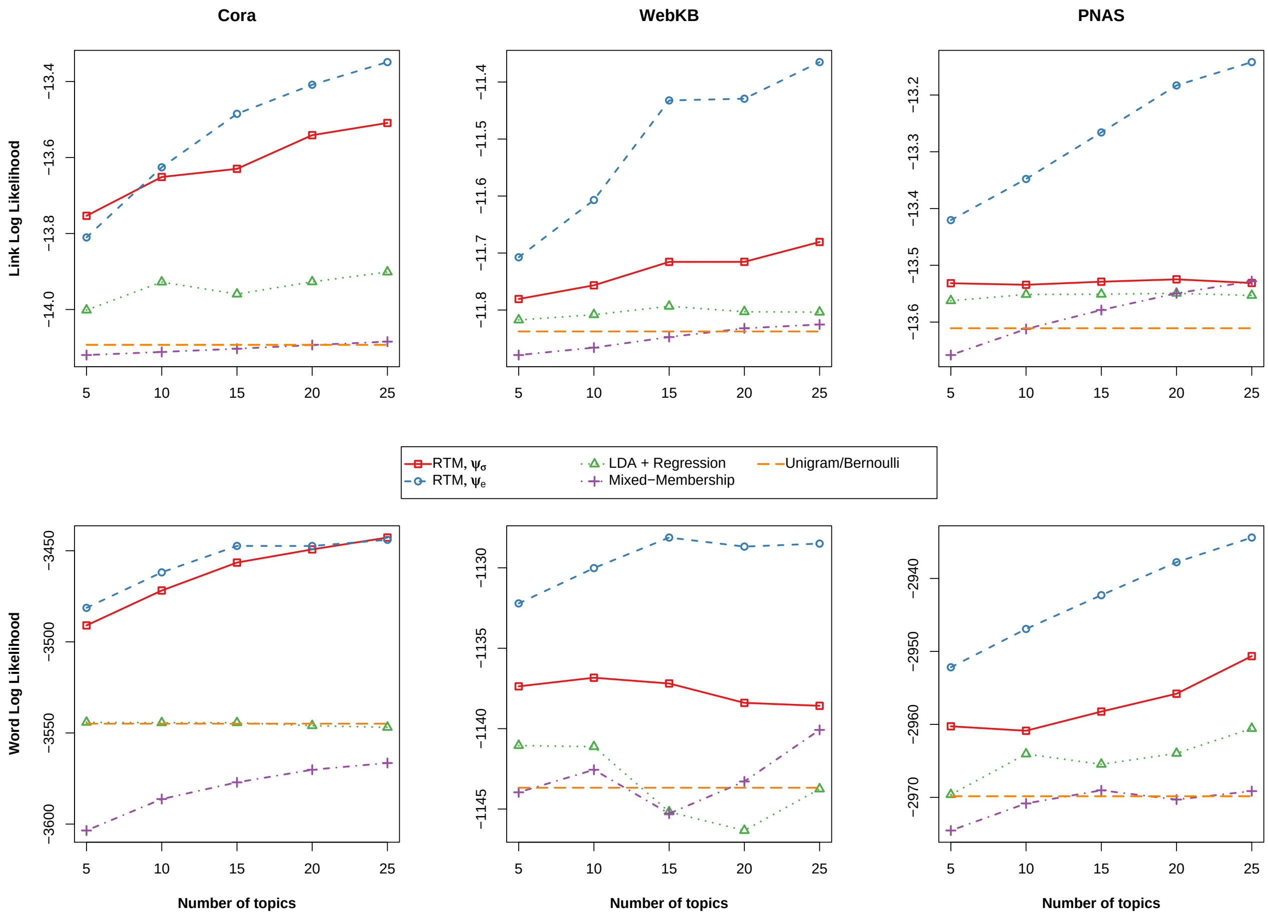


Figure 3: Average held-out predictive link log likelihood (top) and word log likelihood (bottom) as a function of the number of topics. For all three corpora, RTMs outperform baseline unigram, LDA, and "Mixed-Membership," which is the model of Nallapati et al. (2008).

ing set, and other topics to explain the links observed in the training set. Therefore, it cannot use word observations to predict links.

In predicting words, the two variants of the RTM again outperform all of the alternative models (see Figure 3, bottom row). This is because the RTM uses link information to influence the predictive distribution of words. In contrast, the predictions of LDA + Regression are similar to the Baseline. The predictions of the Mixed-Membership model are rarely higher than Baseline, and often lower.

Automatic link suggestion A natural real-world application of link prediction is to suggest links to a user based on the text of a document. One might suggest citations for an abstract or friends for a user in a social network.

Table 1 illustrates suggested citations using RTM (ψ_e) and LDA + Regression as predictive models. These suggestions were computed from a model trained on one of the folds of the *Cora* data. The top results illustrate suggested links for "Markov chain Monte Carlo convergence diagnostics: A comparative review," which occurs in this fold's training set. The bottom results illustrate suggested links for "Competitive environments evolve better solutions for complex tasks," which is in the test set.

RTM outperforms LDA + Regression in being able to identify more true connections. For the first document, RTM finds 3 of the connected documents versus 1 for LDA + Regression. For the second document, RTM finds 3 while LDA + Regression does not find any. This qualitative behavior is borne out quantitatively over the entire corpus. Considering the precision of the first 20 documents retrieved by the models, RTM improves precision over LDA + Regression by 80%. (Twenty is a reasonable number of documents for a user to examine.)

While both models found several connections which were not observed in the data, those found by the RTM are qualitatively different. In the first document, both sets of suggested links are about Markov chain Monte Carlo. However, the RTM finds more documents relating specifically to convergence and stationary behavior of Monte Carlo methods. LDA + Regression finds connections to documents in the milieu of MCMC, but many are only indirectly related

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo	RTM (ψ_e)
Rates of convergence of the Hastings and Metropolis algorithms	
Possible biases induced by MCMC convergence diagnostics	
Bounding convergence time of the Gibbs sampler in Bayesian image restoration	
Self regenerative Markov chain Monte Carlo	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation	
Diagnosing convergence of Markov chain Monte Carlo algorithms	
Exact Bound for the Convergence of Metropolis Chains	**LDA + Regression**
Self regenerative Markov chain Monte Carlo	
Minorization conditions and convergence rates for Markov chain Monte Carlo	
Gibbs-markov models	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models	
Mediating instrumental variables	
A qualitative framework for probabilistic inference	
Adaptation for Self Regenerative MCMC	

Competitive environments evolve better solutions for complex tasks	
Coevolving High Level Representations	RTM (ψ_e)
A Survey of Evolutionary Strategies	
Genetic Algorithms in Search, Optimization and Machine Learning	
Strongly typed genetic programming in evolving cooperation strategies	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems	
Evolutionary Module Acquisition	
An Empirical Investigation of Multi-Parent Recombination Operators in Evolution Strategies	
A New Algorithm for DNA Sequence Assembly	**LDA + Regression**
Identification of protein coding regions in genomic DNA	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems	
A genetic algorithm for passive management	
The Performance of a Genetic Algorithm on a Chaotic Objective Function	
Adaptive global optimization with local search	
Mutation rates as adaptations	

Table 1: Top eight link predictions made by RTM (ψ_e) and LDA + Regression for two documents (italicized) from *Cora*. The models were trained with 10 topics. Boldfaced titles indicate actual documents cited by or citing each document. Over the whole corpus, RTM improves precision over LDA + Regression by 80% when evaluated on the first 20 documents retrieved.

to the input document. The RTM is able to capture that the notion of "convergence" is an important predictor for citations, and has adjusted the topic distribution and predictors correspondingly. For the second document, the documents found by the RTM are also of a different nature than those found by LDA + Regression. All of the documents suggested by RTM relate to genetic algorithms. LDA + Regression, however, suggests some documents which are about genomics. By relying only on words, LDA + Regression conflates two "genetic" topics which are similar in vocabulary but different in citation structure. In contrast, the RTM partitions the latent space differently, recognizing that papers about DNA sequencing are unlikely to cite papers about genetic algorithms, and vice versa. It is better able to capture the joint distribution of words and links.

5 RELATED WORK AND DISCUSSION

The RTM builds on previous research in statistics and machine learning. Many models have been developed to explain network link structure (Wasserman and Pattison 1996; Newman 2002) and extensions which incorporate node attributes have been proposed (Getoor et al. 2001; Taskar et al. 2004). However, these models are not latent space approaches and therefore cannot provide the benefits of dimensionality reduction and produce the interpretable clusters of nodes useful for understanding community structure.

The RTM, in contrast, is a latent space approach which can provide meaningful clusterings of both nodes *and* attributes. Several latent space models for modeling network structure have been proposed (Kemp et al. 2004; Hoff et al. 2002; Hofman and Wiggins 2007; Airoldi et al. 2008); though powerful, these models only account for links in the data and cannot model node attributes as well.

Because the RTM jointly models node attributes and link structure, it can make predictions about one given the other. Previous work tends to explore one or the other of these two prediction problems. Some previous work uses link structure to make attribute predictions (Chakrabarti et al. 1998; Kleinberg 1999), including several topic models (Dietz et al. 2007; McCallum et al. 2005; Wang et al. 2005). However, none of these methods can make predictions about links given words.

In addition to being able to make predictions about links given words and words given links, the RTM is able to do so for *new* documents—documents outside of training data. Approaches which generate document links through topic models (Nallapati and Cohen 2008; Cohn and Hofmann 2001; Sinkkonen et al. 2008; Gruber et al. 2008) treat links as discrete "terms" from a separate vocabulary. This encodes the observed training data into the model, which cannot be generalized to observations outside of it. Link and word predictions for new documents, of the kind we evaluate in Section 4, are ill-defined in these models.

Closest to the RTM is recent work by Nallapati et al. (2008) and Mei et al. (2008), which attempts to address these issues by extending the mixed-membership stochastic block model (Airoldi et al. 2008) to include word attributes. Because of their underlying exchangeability assumptions, these models allow for the links to be explained by some topics and the words to be explained by others. This hinders their predictions when using information about words to predict link structure and vice versa. In contrast, the RTM enforces the constraint that topics be used to explain *both* words and links. We showed in Section 4 that the RTM outperforms such models on these tasks.

The RTM is a new probabilistic generative model of documents and links between them. The RTM is used to analyze linked corpora such as citation networks, linked web pages, and social networks with user profiles. We have demonstrated qualitatively and quantitatively that the RTM provides an effective and useful mechanism for analyzing and using such data. It significantly improves on previous models, integrating both node-specific information and link structure to give better predictions.

Acknowledgements

David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, and grants from Google and Microsoft.

References

E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, pages 1981 – 2014, September 2008.

D. Blei and M. Jordan. Modeling annotated data. *Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003.

D. M. Blei and J. D. McAuliffe. Supervised topic models. *Neural Information Processsing Systems*, Aug 2007.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Arxiv preprint arXiv:0712.2526*, Jan 2007.

S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext classification using hyperlinks. *Proc. ACM SIGMOD*, 1998.

D. Cohn and T. Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems 13*, 2001.

M. Craven, D. DiPasquo, D. Freitag, and A. McCallum. Learning to extract symbolic knowledge from the world wide web. *Proc. AAAI*, 1998.

L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. *Proc. ICML*, 2007.

L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. *Proc. ICML*, 2001.

A. Gruber, M. Rosen-Zvi, and Y. Weiss. Latent topic models for hypertext. *Uncertainty in Artificial Intelligence*, May 2008.

P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 2002.

J. Hofman and C. Wiggins. A Bayesian approach to network modularity. *eprint arXiv: 0709.3512*, 2007.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Oct 1999.

C. Kemp, T. Griffiths, and J. Tenenbaum. Discovering latent classes in relational data. *MIT AI Memo 2004-019*, 2004.

J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 1999.

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

A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, 2005.

Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. *WWW '08: Proceeding of the 17th international conference on World Wide Web*, Apr 2008.

R. Nallapati and W. Cohen. Link-pLSA-LDA: A new unsupervised model for topics and influence of blogs. *ICWSM*, 2008.

R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

M. Newman. The structure and function of networks. *Computer Physics Communications*, 2002.

J. Sinkkonen, J. Aukia, and S. Kaski. Component models for large networks. *arXiv*, stat.ML, Mar 2008.

B. Taskar, M. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. *NIPS*, 2004.

X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and text. *Proceedings of the 3rd international workshop on Link discovery*, 2005.

S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika*, 1996.