A Discriminative Topic Model using Document Network Structure

Weiwei Yang
Computer Science
University of Maryland
College Park, MD
wwyang@cs.umd.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO
Jordan.Boyd.Graber@colorado.edu

Philip Resnik
Linguistics and UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

Abstract

Document collections often have links between documents—citations, hyperlinks, or revisions—and which links are added is often based on topical similarity. To model these intuitions, we introduce a new topic model for documents situated within a network structure, integrating latent blocks of documents with a max-margin learning criterion for link prediction using topic- and word-level features. Experiments on a scientific paper dataset and collection of webpages show that, by more robustly exploiting the rich link structure within a document network, our model improves link prediction, topic quality, and block distributions.

1 Introduction

Documents often appear within a network structure: social media mentions, retweets, and follower relationships; Web pages by hyperlinks; scientific papers by citations. Network structure interacts with the topics in the text, in that documents linked in a network are more likely to have similar topic distributions. For instance, a citation link between two papers suggests that they are about a similar field, and a mentioning link between two social media users often indicates common interests. Conversely, documents' similar topic distributions can suggest links between them. For example, topic model (Blei et al., 2003, LDA) and block detection papers (Holland et al., 1983) are relevant to our research, so we cite them. Similarly, if a social media user A finds another user B with shared interests, then A is more likely to follow B.

Our approach is part of a natural progression of network modeling in which models integrate more information in more sophisticated ways. Some past methods only consider the network itself (Kim and Leskovec, 2012; Liben-Nowell and Kleinberg, 2007), which loses the rich information in text. In other cases, methods take both links and text into account (Chaturvedi et al., 2012), but they are modeled separately, not jointly, limiting the model's ability to capture interactions between the two. The *relational topic model* (Chang and Blei, 2010, RTM) goes further, jointly modeling topics and links, but it considers only pairwise document relationships, failing to capture network structure at the level of groups or *blocks* of documents.

We propose a new joint model that makes fuller use of the rich link structure within a document network. Specifically, our model embeds the *weighted stochastic block model* (Aicher et al., 2014, WSBM) to identify blocks in which documents are densely connected. WSBM basically categorizes each item in a network probabilistically as belonging to one of L blocks, by reviewing its connections with each block. Our model can be viewed as a principled probabilistic extension of Yang et al. (2015), who identify blocks in a document network deterministically as *strongly connected components* (SCC). Like them, we assign a distinct Dirichlet prior to each block to capture its topical commonalities. Jointly, a linear regression model with a discriminative, max-margin objective function (Zhu et al., 2012; Zhu et al., 2014) is trained to reconstruct the links, taking into account the features of documents' topic and word distributions (Nguyen et al., 2013), block assignments, and inter-block link rates.

We validate our approach on a scientific paper abstract dataset and collection of webpages, with citation links and hyperlinks respectively, to predict links among previously unseen documents and from those new documents to training documents. Embedding the WSBM in a network/topic

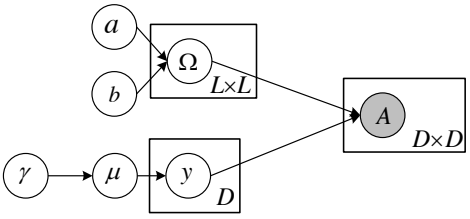


Figure 1: Weighted Stochastic Block Model

model leads to substantial improvements in link prediction over previous models; it also improves block detection and topic interpretability. The key advantage in embedding WSBM is its flexibility and robustness in the face of noisy links. Our results also lend additional support for using max-margin learning for a "downstream" supervised topic model (McAuliffe and Blei, 2008), and that predictions from lexical as well as topic features improves performance (Nguyen et al., 2013).

The rest of this paper is organized as follows. Section 2 introduces two previous link-modeling methods, WSBM and RTM. Section 3 presents our methods to incorporate block priors in topic modeling and include various features in link prediction, as well as the aggregated discriminative topic model whose posterior inference is introduced in Section 4. In Section 5 we show how our model can improve link prediction and (often) improve topic coherence.

2 Dealing with Links

2.1 Weighted Stochastic Block Model

WSBM (Aicher et al., 2014) is a generalized *stochastic block model* (Holland et al., 1983; Wang and Wong, 1987, SBM) and predicts non-negative integer-weight links, instead of binary-weight links. A block is a collection of documents which are densely connected with each other but sparsely connected with documents in other blocks. WSBM assumes that a document belongs to exactly one block. A link connecting two documents in blocks l and l' has a weight generated from a Poisson distribution with parameters $\Omega_{l,l'}$ which has a Gamma prior with parameters a and b, as Figure 1 shows.

The whole generative process is:

1. For each pair of blocks $(l, l') \in \{1, \dots, L\}^2$
   (a) Draw inter-block link rate $\Omega_{l,l'} \sim \text{Gamma}(a, b)$
2. Draw block distribution $\boldsymbol{\mu} \sim \text{Dir}(\gamma)$
3. For each document $d \in \{1, \dots, D\}$
   (a) Draw block assignment $y_d \sim \text{Mult}(\boldsymbol{\mu})$
4. For each link $(d, d') \in \{1, \dots, D\}^2$
   (a) Draw link weight $A_{d,d'} \sim \text{Poisson}(\Omega_{y_d, y_{d'}})$

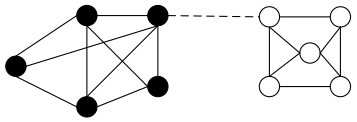


Figure 2: SCC can be distracted by spurious links connecting two groups, while WSBM maintains the distinction.

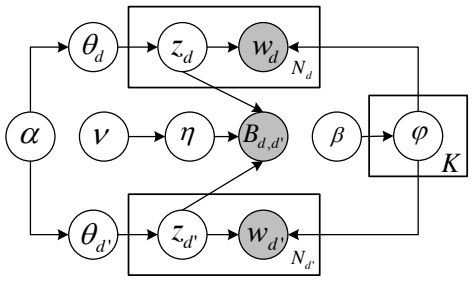


Figure 3: A Two-document Segment of RTM

WSBM is a probabilistic block detection algorithm and more robust than some deterministic algorithms like SCC, which is vulnerable to noisy links. For instance, we would intuitively say Figure 2 has two blocks—as denoted by coloring—whether or not the dashed link exists. If the dashed link does not exist, both WSBM and SCC can identify two blocks. However, if the dashed link does exist, SCC will return only one big block that contains all nodes, while WSBM still keeps the nodes in two reasonable blocks.

2.2 Relational Topic Model

RTM (Chang and Blei, 2010) is a downstream model that generates documents and links simultaneously (Figure 3). Its generative process is:

1. For each topic $k \in \{1, \dots, K\}$
   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dir}(\beta)$
   (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$
2. For each document $d \in \{1, \dots, D\}$
   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dir}(\alpha)$
   (b) For each token $t_{d,n}$ in document d
      i. Draw topic assignment $z_{d,n} \sim \text{Mult}(\boldsymbol{\theta_d})$
      ii. Draw word $w_{d,n} \sim \text{Mult}(\boldsymbol{\phi}_{\boldsymbol{z_{d,n}}})$
3. For each explicit link (d, d')
   (a) Draw link weight $B_{d,d'} \sim \Psi(\cdot \mid \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{\eta})$

In the inference process, the updating of topic assignments is guided by links so that linked documents are more likely to have similar topic distributions. Meanwhile, the linear regression (whose

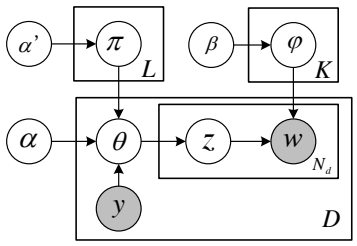


Figure 4: Graphical Model of BP-LDA

output is fed into link probability function Ψ) is updated to maximize the network likelihood using current topic assignments.

3 Discriminative Topic Model with Block Prior and Various Features

Our model is able to identify blocks from the network with an embedded WSBM, extract topic patterns of each block as prior knowledge, and use all this information to reconstruct the links.

3.1 LDA with Block Priors (BP-LDA)

As argued in the introduction, linked documents are likely to have similar topic distributions, which can be generalized to documents in the same block. Inspired by this intuition and the block assignment we obtain in the previous section, we want to extract some prior knowledge from these blocks. Thus we propose an LDA with block priors, hence BP-LDA, as shown in Figure 4, which has the following generative process:

1. For each topic $k \in \{1, \dots, K\}$
   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dir}(\beta)$
2. For each block $l \in \{1, \dots, L\}$
   (a) Draw topic distribution $\boldsymbol{\pi_l} \sim \text{Dir}(\alpha')$
3. For each document $d \in \{1, \dots, D\}$
   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dir}(\alpha \boldsymbol{\pi_{y_d}})$
   (b) For each token $t_{d,n}$ in document d
      i. Draw topic assignment $z_{d,n} \sim \text{Mult}(\boldsymbol{\theta_d})$
      ii. Draw word $w_{d,n} \sim \text{Mult}(\boldsymbol{\phi_{z_{d,n}}})$

Unlike conventional LDA, which uses an uninformative topic prior, BP-LDA puts a Dirichlet prior $\boldsymbol{\pi}$ on each block to capture the block's topic distribution and use it as an informative prior when drawing each document's topic distribution. In other words, a document's topic distribution—i.e., what the document is about—is not just informed by the words present in the document but the broader context of its network neighborhood.

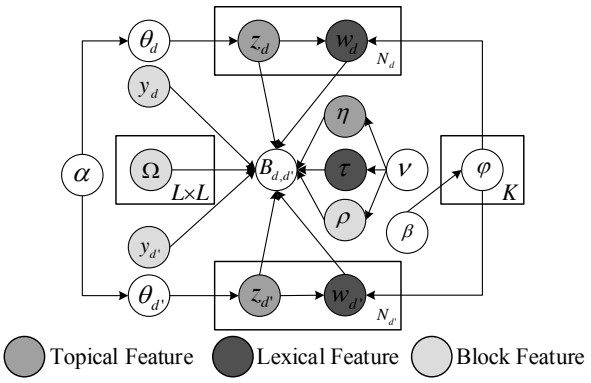


Figure 5: A two-document segment of VF-RTM. Various features are denoted by grayscale. $B_{d,d'}$ is observed, but we keep it in white background to avoid confusion.

3.2 RTM with Various Features (VF-RTM)

Building on Chang and Blei (2010), we want to generate the links between documents based on various features, hence VF-RTM. In addition to topic distributions, VF-RTM also includes documents' word distributions (Nguyen et al., 2013) and the link rate of two documents' assigned blocks, with the intent that these additional features improve link generation. VF-RTM involves the relationship between a pair of documents, so it is difficult to show the whole model; therefore Figure 5 illustrates with a two-document segment. The generative process is:

1. For each pair of blocks $(l, l') \in \{1, \dots, L\}^2$
   (a) Draw block regression parameter $\rho_{l,l'} \sim \mathcal{N}(0, \nu^2)$
2. For each topic $k \in \{1, \dots, K\}$
   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dir}(\beta)$
   (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$
3. For each word $v \in \{1, \dots, V\}$
   (a) Draw lexical regression parameter $\tau_v \sim \mathcal{N}(0, \nu^2)$
4. For each document $d \in \{1, \dots, D\}$
   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dir}(\alpha)$
   (b) For each token $t_{d,n}$ in document d
      i. Draw topic assignment $z_{d,n} \sim \text{Mult}(\boldsymbol{\theta_d})$
      ii. Draw word $w_{d,n} \sim \text{Mult}(\boldsymbol{\phi_{z_{d,n}}})$
5. For each explicit link (d, d')
   (a) Draw link weight
   $B_{d,d'} \sim \Psi(\cdot \mid y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho})$

Links are generated by a link probability function Ψ which takes the regression value $R_{d,d'}$ of documents d and d' as an argument. Assuming documents d and d' belong to blocks l and l' respectively, $R_{d,d'}$ is

$$R_{d,d'} = \boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'}) + \rho_{l,l'}\Omega_{l,l'}, \quad (1)$$

where $\overline{\boldsymbol{z}}_d$ is a K-length vector with each element $\overline{z}_{d,k} = \frac{1}{N_d}\sum_n \mathbb{1}(z_{d,n} = k)$; $\overline{\boldsymbol{w}}_d$ is a V-length vector with each element $\overline{w}_{d,v} = \frac{1}{N_d}\sum_n \mathbb{1}(w_{d,n} = v)$; $\circ$ denotes the Hadamard (element-wise) product;[1] $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ are the weight vectors and matrix for topic-based, lexical-based and rate-based predictions, respectively.

A common choice of Ψ is a sigmoid (Chang and Blei, 2010). However, we instead use hinge loss so that VF-RTM can use the max-margin principle, making more effective use of side information when inferring topic assignments (Zhu et al., 2012). Using hinge loss, the probability that documents d and d' are linked is

$$\Pr(B_{d,d'}) = \exp(-2\max(0, \zeta_{d,d'})), \quad (2)$$

where $\zeta_{d,d'} = 1 - B_{d,d'}R_{d,d'}$. Positive and negative link weights are denoted by 1 and -1, respectively, in contrast to sigmoid loss.

3.3 Aggregated Model

Finally, we put all the pieces together and propose LBH-RTM: RTM with lexical weights (L), block priors (B), and hinge loss (H). Its graphical model is given in Figure 6.

1. For each pair of blocks $(l, l') \in \{1, \dots, L\}^2$
   (a) Draw inter-block link rate $\Omega_{l,l'} \sim \text{Gamma}(a, b)$
   (b) Draw block regression parameter $\rho_{l,l'} \sim \mathcal{N}(0, \nu^2)$
2. Draw block distribution $\boldsymbol{\mu} \sim \text{Dir}(\gamma)$
3. For each block $l \in \{1, \dots, L\}$
   (a) Draw topic distribution $\boldsymbol{\pi_l} \sim \text{Dir}(\alpha')$
4. For each topic $k \in \{1, \dots, K\}$
   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dir}(\beta)$
   (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$
5. For each word $v \in \{1, \dots, V\}$
   (a) Draw lexical regression parameter $\tau_v \sim \mathcal{N}(0, \nu^2)$
6. For each document $d \in \{1, \dots, D\}$
   (a) Draw block assignment $y_d \sim \text{Mult}(\boldsymbol{\mu})$
   (b) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dir}(\alpha\boldsymbol{\pi}_{\boldsymbol{y_d}})$
   (c) For each token $t_{d,n}$ in document d
      i. Draw topic assignment $z_{d,n} \sim \text{Mult}(\boldsymbol{\theta_d})$
      ii. Draw word $w_{d,n} \sim \text{Mult}(\boldsymbol{\phi}_{\boldsymbol{z_{d,n}}})$
7. For each link $(d, d') \in \{1, \dots, D\}^2$
   (a) Draw link weight $A_{d,d'} \sim \text{Poisson}(\Omega_{y_d, y_{d'}})$
8. For each explicit link (d, d')
   (a) Draw link weight $B_{d,d'} \sim \Psi(\cdot \mid y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho})$

$\boldsymbol{A}$ and $\boldsymbol{B}$ are assumed independent in the model, but they can be derived from the same set of links in practice.

[1] As Chang and Blei (2010) point out, the Hadamard product is able to capture similarity between hidden topic representations of two documents.

Algorithm 1 Sampling Process

1: Set $\boldsymbol{\lambda} = 1$ and initialize topic assignments
2: **for** $m = 1$ to M **do**
3: Optimize $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ using L-BFGS
4: **for** $d = 1$ to D **do**
5: Draw block assignment y_d
6: **for** each token n **do**
7: Draw a topic assignment $z_{d,n}$
8: **end for**
9: **for** each explicit link (d, d') **do**
10: Draw $\lambda^{-1}_{d,d'}$ (and then $\lambda_{d,d'}$)
11: **end for**
12: **end for**
13: **end for**

Link set $\boldsymbol{A}$ is primarily used to find blocks, so it treats all links *deterministically*. In other words, the links observed in the input are considered explicit positive links, while the unobserved links are considered explicit negative links, in contrast to the implicit links in $\boldsymbol{B}$.

In terms of link set $\boldsymbol{B}$, while it adopts all explicit positive links from the input, it *does not deny* the existence of unobserved links, or implicit negative links. Thus $\boldsymbol{B}$ consists of only explicit positive links. However, to avoid overfitting, we sample some implicit links and add them to $\boldsymbol{B}$ as explicit negative links.

4 Posterior Inference

Posterior inference (Algorithm 1) consists of the sampling of topic and block assignments and the optimization of weight vectors and matrix.[2] We add an auxiliary variable $\boldsymbol{\lambda}$ for hinge loss (see Section 4.2), so the updating of $\boldsymbol{\lambda}$ is not necessary when using sigmoid loss.

The sampling procedure is an iterative process after initialization (Line 1). In each iteration, we first optimize the weight vectors and matrix (Line 3) before updating documents' block assignments (Line 5) and topic assignments (Line 7). When using hinge loss, the auxiliary variable $\boldsymbol{\lambda}$ for every explicit link needs to be updated (Line 10).

4.1 Sampling Block Assignments

Block assignment sampling is done by Gibbs sampling, using the block assignments and links in $\boldsymbol{A}$

[2] More details about sampling procedures and equations in this section (including the sampling and optimization equations using sigmoid loss) are available in the supplementary material.

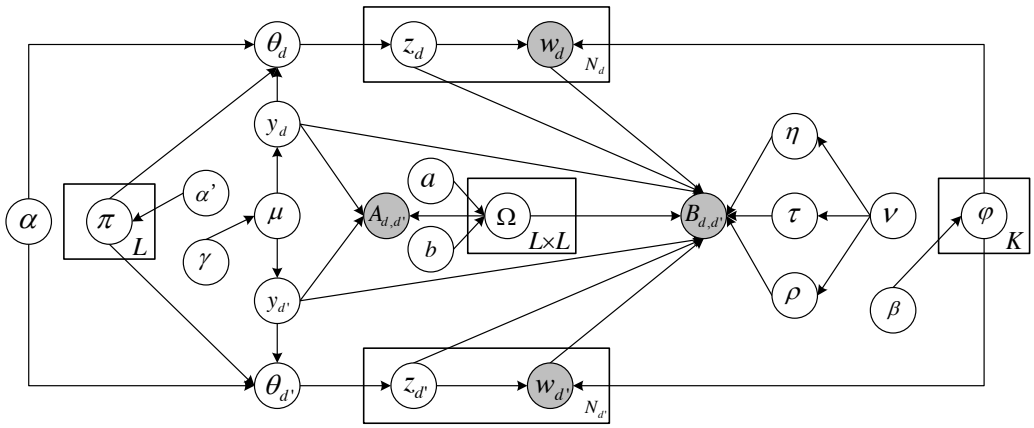


Figure 6: The graphical model of LBH-RTM for two documents, in which a weighted stochastic block model is embedded (γ, μ, y, a, b, $\boldsymbol{\Omega}$, and $\boldsymbol{A}$). Each document's topic distribution has an informative prior $\boldsymbol{\pi}$. The model predicts links between documents ($\boldsymbol{B}$) based on topics ($\boldsymbol{z}$), words ($\boldsymbol{w}$), and inter-block link rates ($\boldsymbol{\Omega}$), using a max-margin objective.

excluding document d and its related links.[3] The probability that d is assigned to block l is

$$\Pr(y_d = l \,|\, \boldsymbol{A}_{-d}, \boldsymbol{y}_{-d}, a, b, \gamma) \propto \left(N_l^{-d} + \gamma\right) \times \prod_{l'} \frac{\left(S_e^{-d}(l,l') + b\right)^{S_w^{-d}(l,l')+a}}{\left(S_e^{-d}(l,l') + b + S_e(d,l')\right)^{S_w^{-d}(l,l')+a+S_w(d,l')}} \prod_{i=0}^{S_w(d,l')-1} \left(S_w^{-d}(l,l') + a + i\right), \quad (3)$$

where N_l is the number of documents assigned to block l; $^{-d}$ denotes that the count excludes document d; $S_w(d,l)$ and $S_w(l,l')$ are the sums of link weights from document d to block l and from block l to block l', respectively:

$$S_w(d,l) = \sum_{d':y_{d'}=l} A_{d,d'} \quad (4)$$

$$S_w(l,l') = \sum_{d:y_d=l} S_w(d,l'). \quad (5)$$

$S_e(d,l)$ is the number of possible links from document d to l (i.e., assuming document d connects to every document in block l), which equals N_l. The number of possible links from block l to l' is $S_e(l,l')$ (i.e., assuming every document in block l connects to every document in block l'):

$$S_e(l,l') = \begin{cases} N_l \times N_{l'} & l \neq l' \\ \frac{1}{2} N_l(N_l - 1) & l = l'. \end{cases} \quad (6)$$

If we rearrange the terms of Equation 3 and put the terms which have $S_w(d,l')$ together, the value of $S_w(d,l')$ increases (i.e., document d is more densely connected with documents in block l'), the probability of assigning d to block l decreases exponentially. Thus if d is more densely connected with block l and sparsely connected with other blocks, it is more likely to be assigned to block l.

4.2 Sampling Topic Assignments

Following Polson and Scott (2011), by introducing an auxiliary variable $\lambda_{d,d'}$, the conditional probability of assigning $t_{d,n}$, the n-th token in document d, to topic k is

$$\Pr(z_{d,n} = k \,|\, \boldsymbol{z}_{-d,n}, \boldsymbol{w}_{-d,n}, w_{d,n} = v, y_d = l) \propto \left(N_{d,k}^{-d,n} + \alpha \pi_{l,k}^{-d,n}\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} \prod_{d'} \exp\left(-\frac{(\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), \quad (7)$$

where $N_{d,k}$ is the number of tokens in document d that are assigned to topic k; $N_{k,v}$ denotes the count of word v assigned to topic k; Marginal counts are denoted by $\cdot$; $^{-d,n}$ denotes that the count excludes $t_{d,n}$; d' denotes all documents that have explicit links with document d. The block topic prior $\pi_{l,k}^{-d,n}$ is estimated based on the maximal path assumption (Cowans, 2006; Wallach, 2008):

$$\pi_{l,k}^{-d,n} = \frac{\sum_{d':y_{d'}=l} N_{d',k}^{-d,n} + \alpha'}{\sum_{d':y_{d'}=l} N_{d',\cdot}^{-d,n} + K\alpha'}. \quad (8)$$

the link prediction argument $\zeta_{d,d'}$ is

$$\zeta_{d,d'} = 1 - B_{d,d'}\left(\frac{\eta_k}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} + R_{d,d'}^{-d,n}\right). \quad (9)$$

[3] These equations deal with undirected edges, but they can be adapted for directed edges. See supplementary material.

where

$$R_{d,d'}^{-d,n} = \sum_{k=1}^{K} \eta_k \frac{N_{d,k}^{-d,n}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} + \sum_{v=1}^{V} \tau_v \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d, y_{d'}} \Omega_{y_d, y_{d'}}. \quad (10)$$

Looking at the first term of Equation 7, the probability of assigning $t_{d,n}$ to topic k depends not only on its own topic distribution, but also the topic distribution of the block it belongs to. The links also matter: Equation 9 gives us the intuition that a topic which could increase the likelihood of links is more likely to be selected, which forms an interaction between topics and the link graph—the links are guiding the topic sampling while updating topic assignments is maximizing the likelihood of the link graph.

4.3 Parameter Optimization

While topic assignments are updated iteratively, the weight vectors and matrix $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ are optimized in each global iteration over the whole corpus using L-BFGS (Liu and Nocedal, 1989). It takes the likelihood of generating $\boldsymbol{B}$ using $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and current topic and block assignments as the objective function, and optimizes it using the partial derivatives with respect to every weight vector/matrix element.

The log likelihood of $\boldsymbol{B}$ using hinge loss is

$$\mathcal{L}(\boldsymbol{B}) \propto -\sum_{d,d'} \frac{R_{d,d'}^2 - 2(1+\lambda_{d,d'}) B_{d,d'} R_{d,d'}}{2\lambda_{d,d'}} - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L} \sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2}. \quad (11)$$

We also need to update the auxiliary variable $\lambda_{d,d'}$. Since the likelihood of $\lambda_{d,d'}$ follows a generalized inverse Gaussian distribution $\mathrm{GIG}\left(\lambda_{d,d'}; \frac{1}{2}, 1, \zeta_{d,d'}^2\right)$, we sample its reciprocal $\lambda_{d,d'}^{-1}$ from an inverse Gaussian distribution as

$$\Pr\left(\lambda_{d,d'}^{-1} \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) = \mathrm{IG}\left(\lambda_{d,d'}^{-1}; \frac{1}{|\zeta_{d,d'}|}, 1\right). \quad (12)$$

5 Experimental Results

We evaluate using the two datasets. The first one is CORA dataset (McCallum et al., 2000). After removing stopwords and words that appear in fewer than ten documents, as well as documents with no words or links, our vocabulary has 1,240 unique words. The corpus has 2,362 computer science paper abstracts with 4,231 citation links.

Model	PLR	
	CORA	WEBKB
RTM (Chang and Blei, 2010)	419.33	141.65
LCH-RTM (Yang et al., 2015)	459.55	150.32
BS-RTM	391.88	127.25
LBS-RTM	383.25	125.41
LBH-RTM	**360.38**	**111.79**

Table 1: Predictive Link Rank Results

The second dataset is WEBKB. It is already preprocessed and has 1,703 unique words in vocabulary. The corpus has 877 web pages with 1,608 hyperlinks.

We treat all links as undirected. Both datasets are split into 5 folds, each further split into a development and test set with approximately the same size when used for evaluation.

5.1 Link Prediction Results

In this section, we evaluate LBH-RTM and its variations on link prediction tasks using *predictive link rank* (PLR). A document's PLR is the average rank of the documents to which it has explicit positive links, among all documents, so lower PLR is better.

Following the experiment setup in Chang and Blei (2010), we train the models on the training set and predict citation links within held-out documents as well as from held-out documents to training documents. We tune two important parameters—α and negative edge ratio (the ratio of the number of sampled negative links to the number of explicit positive links)—on the development set and apply the trained model which performs the best on the development set to the test set.[4] The cross validation results are given in Table 1, where models are differently equipped with lexical weights (L), WSBM prior (B), SCC prior (C), hinge loss (H), and sigmoid loss (S).[5] Link prediction generally improves with incremental application of prior knowledge and more sophisticated learning techniques.

The embedded WSBM brings around 6.5% and 10.2% improvement over RTM in PLR on the

[4] We also tune the number of blocks for embedded WSBM and set it to 35 (CORA) and 20 (WEBKB). The block topic priors are not applied on unseen documents, since we don't have available links.

[5] The values of RTM are different from the result reported by Chang and Blei (2010), because we re-preprocessed the CORA dataset and used different parameters.

CORA and WEBKB datasets, respectively. This indicates that the blocks identified by WSBM are reasonable and consistent with reality. The lexical weights also help link prediction (LBS-RTM), though less for BS-RTM. This is understandable since word distributions are much sparser and do not make as significant a contribution as topic distributions. Finally, hinge loss improves PLR substantially (LBH-RTM), about 14.1% and 21.1% improvement over RTM on the CORA and WEBKB datasets respectively, demonstrating the effectiveness of max-margin learning.

The only difference between LCH-RTM and LBH-RTM is the block detection algorithm. However, their link prediction performance is poles apart—LCH-RTM even fails to outperform RTM. This implies that the quality of blocks identified by SCC is not as good as WSBM, which we also illustrate in Section 5.4.

5.2 Illustrative Example

We illustrate our model's behavior qualitatively by looking at two abstracts, Koplon and Sontag (1997) and Albertini and Sontag (1992) from the CORA dataset, designated K and A for short.

Paper K studies the application of Fourier-type activation functions in fully recurrent neural networks. Paper A shows that if two neural networks have equal behaviors as "black boxes", they must have the same number of neurons and the same weights (except sign reversals).

From the titles and abstracts, we can easily find that both of them are about *neural networks* (NN). They both contain words like *neural*, *neuron*, *network*, *recurrent*, *activation*, and *nonlinear*, which corresponds to the topic with words *neural*, *network*, *train*, *learn*, *function*, *recurrent*, etc. There is a citation between K and A. The ranking of this link improves as the model gets more sophisticated (Table 2), except LCH-RTM, which is consistent with our PLR results.

In Figure 7, we also show the proportions of topics that dominate the two documents according to the various models. There are multiple topics dominating K and A according to RTM (Figure 7(a)). As the model gets more sophisticated, the NN topic proportion gets higher. Finally, only the NN topic dominates the two documents when LBH-RTM is applied (Figure 7(e)).

LCH-RTM gives the highest proportion to the NN topic (Figure 7(b)). However, the NN topic splits into two topics and the proportions are not assigned to the same topic, which greatly brings down the link prediction performance. The splitting of the NN topic also happens in other models (Figure 7(a) and 7(d)), but they assign proportions to the same topic(s). Further comparing with LBH-RTM, the blocks detected by SCC are not improving the modeling of topics and links—some documents that should be in two different blocks are assigned to the same one, as we will show in Section 5.4.

Model	Rank of the Link
RTM	1,265
LCH-RTM	1,385
BS-RTM	635
LBS-RTM	132
LBH-RTM	**106**

Table 2: PLR of the citation link between example documents K and A (described in Section 5.2)

Model	FET		LLR	
	CORA	WEBKB	CORA	WEBKB
RTM	0.1330	0.1312	3.001	6.055
LCH-RTM	0.1418	0.1678	3.071	6.577
BS-RTM	0.1415	0.1950	3.033	6.418
LBS-RTM	0.1342	0.1963	2.984	6.212
LBH-RTM	**0.1453**	**0.2628**	**3.105**	**6.669**

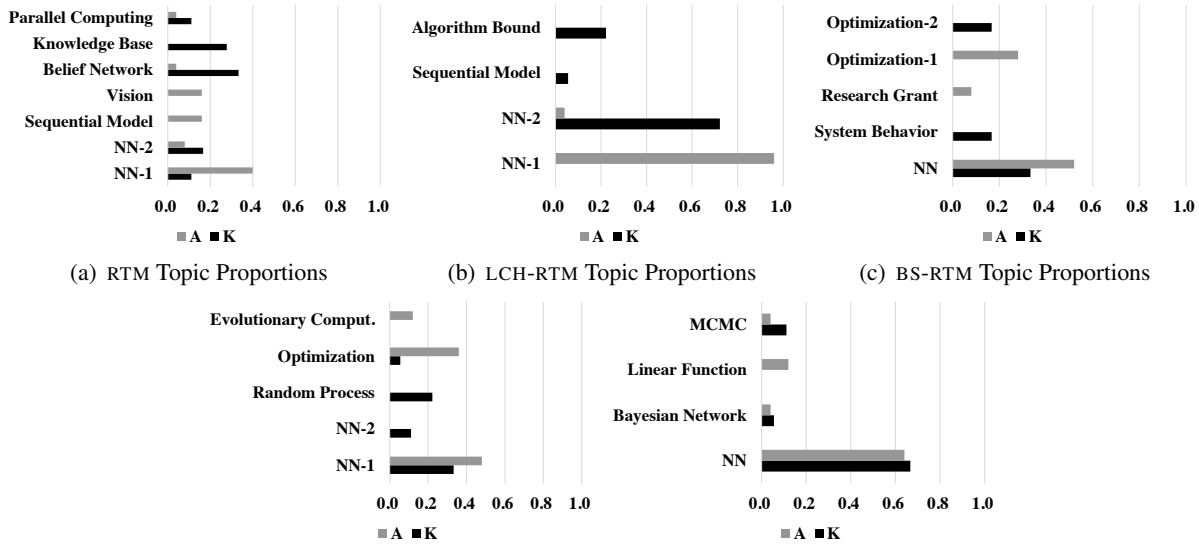
Table 3: Average Association Scores of Topics

5.3 Topic Quality Results

We use an automatic coherence detection method (Lau et al., 2014) to evaluate topic quality. Specifically, for each topic, we pick out the top n words and compute the average association score of each pair of words, based on the held-out documents in development and test sets.

We choose $n = 25$ and use *Fisher's exact test* (Upton, 1992, FET) and *log likelihood ratio* (Moore, 2004, LLR) as the association measures (Table 3). The main advantage of these measures is that they are robust even when the reference corpus is not large.

Coherence improves with WSBM and max-margin learning, but drops a little when adding lexical weights except the FET score on the WEBKB dataset, because lexical weights are intended to improve link prediction performance, not topic quality. Topic quality of LBH-RTM is also better than that of LCH-RTM, suggesting that WSBM benefits topic quality more than SCC.



(a) RTM Topic Proportions (b) LCH-RTM Topic Proportions (c) BS-RTM Topic Proportions

(d) LBS-RTM Topic Proportions (e) LBH-RTM Topic Proportions

Figure 7: Topic proportions given by various models on our two illustrative documents (K and A, described in described in Section 5.2). As the model gets more sophisticated, the NN topic proportion gets higher and finally dominates the two documents when LBH-RTM is applied. Though LCH-RTM gives the highest proportion to the NN topic, it splits the NN topic into two and does not assign the proportions to the same one.

Block	1	2
#Nodes	42	84
#Links in the Block	55	142
#Links across Blocks	2	

Table 4: Statistics of Blocks 1 (learning theory) and 2 (Bayes nets), which are merged in SCC.

5.4 Block Analysis

In this section, we illustrate the effectiveness of the embedded WSBM over SCC.[6] As we have argued, WSBM is able to separate two internally densely-connected blocks even if there are few links connecting them, while SCC tends to merge them in this case. As an example, we focus on two blocks in the CORA dataset identified by WSBM, designated Blocks 1 and 2. Some statistics are given in Table 4. The two blocks are very sparsely connected, but comparatively quite densely connected inside either block. The two blocks' topic distributions also reveal their differences: abstracts in Block 1 mainly focus on learning theory (*learn*, *algorithm*, *bound*, *result*, etc.) and MCMC (*markov*, *chain*, *distribution*, *converge*, etc.). Abstracts in Block 2, however, have higher weights on Bayesian networks (*network*, *model*, *learn*, *bayesian*, etc.) and Bayesian estimation (*estimate*, *bayesian*, *parameter*, *analysis*, etc.), which differs from Block 1's emphasis. Because of the two inter-block links, SCC merges the two blocks into one, which makes the block topic distribution unclear and misleads the sampler. WSBM, on the other hand, keeps the two blocks separate, which generates a high-quality prior for the sampler.

[6] We omit the comparison of WSBM with other models, because this has been done by Aicher et al. (2014). In addition, WSBM is a probabilistic method while SCC is deterministic. They are not comparable quantitatively, so we compare them qualitatively.

6 Related Work

Topic models are widely used in information retrieval (Wei and Croft, 2006), word sense disambiguation (Boyd-Graber et al., 2007), dialogue segmentation (Purver et al., 2006), and collaborative filtering (Marlin, 2003).

Topic models can be extended in either *upstream* or *downstream* way. *Upstream* models generate topics conditioned on supervisory information (Daumé III, 2009; Mimno and McCallum, 2012; Li and Perona, 2005). *Downstream* models, on the contrary, generates topics and supervisory data simultaneously, which turns unsupervised topic models to (semi-)supervised ones. Supervisory data, like labels of documents and links between documents, can be generated from either a maximum likelihood estimation approach (McAuliffe and Blei, 2008; Chang and

Blei, 2010; Boyd-Graber and Resnik, 2010) or a maximum entropy discrimination approach (Zhu et al., 2012; Yang et al., 2015).

In block detection literature, stochastic block model (Holland et al., 1983; Wang and Wong, 1987, SBM) is one of the most basic generative models dealing with binary-weighted edges. SBM assumes that each node belongs to only one block and each link exists with a probability that depends on the block assignments of its connecting nodes. It has been generalized for degree-correction (Karrer and Newman, 2011), bipartite structure (Larremore et al., 2014), and categorial values (Guimerà and Sales-Pardo, 2013), as well as nonnegative integer-weight network (Aicher et al., 2014, WSBM).

Our model combines both topic model and block detection in a unified framework. It takes text, links, and the interaction between text and links into account simultaneously, contrast to the methods that only consider graph structure (Kim and Leskovec, 2012; Liben-Nowell and Kleinberg, 2007) or separate text and links (Chaturvedi et al., 2012).

7 Conclusions and Future Work

We introduce LBH-RTM, a discriminative topic model that jointly models topics and document links, detecting blocks in the document network probabilistically by embedding the weighted stochastic block model, rather via connected-components as in previous models. A separate Dirichlet prior for each block captures its topic preferences, serving as an informed prior when inferring documents' topic distributions. Max-margin learning learns to predict links from documents' topic and word distributions and block assignments.

Our model better captures the connections and content of paper abstracts, as measured by predictive link rank and topic quality. LBH-RTM yields topics with better coherence, though not all techniques contribute to the improvement. We support our quantitative results with qualitative analysis looking at a pair of example documents and at a pair of blocks, highlighting the robustness of embedded WSBM over blocks defined as SCC.

As next steps, we plan to explore model variations to support a wider range of use cases. For example, although we have presented a version of the model defined using undirected binary weight edges in the experiment, it would be straightforward to adapt to model both directed/undirected and binary/nonnegative real weight edges. We are also interested in modeling changing topics and vocabularies (Blei and Lafferty, 2006; Zhai and Boyd-Graber, 2013). In the spirit of treating links probabilistically, we plan to explore application of the model in suggesting links that do not exist but should, for example in discovering missed citations, marking social dynamics (Nguyen et al., 2014), and identifying topically related content in multilingual networks of documents (Hu et al., 2014).

Acknowledgment

This research has been supported in part, under subcontract to Raytheon BBN Technologies, by DARPA award HR0011-15-C-0113. Boyd-Graber is also supported by NSF grants IIS/1320538, IIS/1409287, and NCSE/1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks*.

Francesca Albertini and Eduardo D. Sontag. 1992. For neural networks, function determines form. In *Proceedings of IEEE Conference on Decision and Control*.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the International Conference of Machine Learning*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jonathan Chang and David M. Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*.

Snigdha Chaturvedi, Hal Daumé III, Taesun Moon, and Shashank Srivastava. 2012. A topical graph kernel for link prediction in labeled graphs. In *Proceedings of the International Conference of Machine Learning*.

Philip J. Cowans. 2006. *Probabilistic Document Modelling*. Ph.D. thesis, University of Cambridge.

Hal Daumé III. 2009. Markov random topic fields. In *Proceedings of the Association for Computational Linguistics*.

Roger Guimerà and Marta Sales-Pardo. 2013. A network inference method for large-scale unsupervised identification of novel drug-drug interactions. *PLoS Computational Biology*.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social Networks*.

Yuening Hu, Ke Zhai, Vlad Eidelman, and Jordan Boyd-Graber. 2014. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the Association for Computational Linguistics*.

Brian Karrer and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E*.

Myunghwan Kim and Jure Leskovec. 2012. Latent multi-group membership graph model. In *Proceedings of the International Conference of Machine Learning*.

Renée Koplon and Eduardo D. Sontag. 1997. Using Fourier-neural recurrent networks to fit sequential input/output data. *Neurocomputing*.

Daniel B. Larremore, Aaron Clauset, and Abigail Z. Jacobs. 2014. Efficiently inferring community structure in bipartite networks. *Physical Review E*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the Association for Computational Linguistics*.

Fei-Fei Li and Pietro Perona. 2005. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition*.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*.

Benjamin Marlin. 2003. Modeling user rating profiles for collaborative filtering. In *Proceedings of Advances in Neural Information Processing Systems*.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of Internet portals with machine learning. *Information Retrieval*.

David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of Uncertainty in Artificial Intelligence*.

Robert Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of Empirical Methods in Natural Language Processing*.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah Cai, Jennifer Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*.

Nicholas G. Polson and Steven L. Scott. 2011. Data augmentation for support vector machines. *Bayesian Analysis*.

Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the Association for Computational Linguistics*.

Graham JG Upton. 1992. Fisher's exact test. *Journal of the Royal Statistical Society*.

Hanna M. Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Yuchung J. Wang and George Y. Wong. 1987. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*.

Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2015. Birds of a feather linked together: A discriminative topic model using link-based priors. In *Proceedings of Empirical Methods in Natural Language Processing*.

Ke Zhai and Jordan Boyd-Graber. 2013. Online latent Dirichlet allocation with infinite vocabulary. In *Proceedings of the International Conference of Machine Learning*.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2012. MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research*.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. 2014. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*.