Topic-Link LDA: Joint Models of Topic and Author Community

Yan Liu, Alexandru Niculescu-Mizil {LIUYA, ANICULE}@US.IBM.COM
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Wojciech Gryc WOJCIECH.GRYC@MATHS.OX.AC.UK
Computing Laboratory, Oxford University, Wolfson Building, Parks Road, Oxford OX1 3QD, UK

Abstract

Given a large-scale linked document collection, such as a collection of blog posts or a research literature archive, there are two fundamental problems that have generated a lot of interest in the research community. One is to identify a set of high-level topics covered by the documents in the collection; the other is to uncover and analyze the social network of the authors of the documents. So far these problems have been viewed as separate problems and considered independently from each other. In this paper we argue that these two problems are in fact inter-dependent and should be addressed together. We develop a Bayesian hierarchical approach that performs topic modeling and author community discovery in one unified framework. The effectiveness of our model is demonstrated on two blog data sets in different domains and one research paper citation data from CiteSeer.

1. Introduction

When analyzing a collection of linked documents, we face two fundamental challenges, i.e. monitoring the topics covered by the documents, and uncovering the social network between the authors of the documents. While both topic modeling and social network analysis have received considerable attention from the research community, until now these two tasks have usually been treated independently. For example, there has been significant progress on graphical model approaches for topic modeling, which aims to discover the patterns that reflect the underlying topics form the documents (Blei et al., 2003; Rosen-Zvi et al., 2004; Griffiths & Steyvers, 2004; Blei & Lafferty, 2006; Mei et al., 2008). Simultaneously, many efforts have been devoted to detect the communities from hyperlink via graph mining (Gibson et al., 1998; Chakrabarti & Faloutsos, 2006).

Current solutions to both topic modeling and community discovery have one major drawback: they treat all links (or missing links) between documents the same, which usually is not true in practice. For example, in the case of blog posts, a link between two posts sharing little or no content similarity usually happens when blogger A is a friend of blogger B. This type of link should not be treated the same as those links between posts with strong content similarity, in which case the two bloggers simply discuss same topics without necessarily being part of the same community. Furthermore, the missing links between two posts with strong content similarity indicates more information (i.e. most likely the two bloggers do not know each other) than those missing links with no content similarity.

In this paper, we are interested in jointly modeling the document topics and the social network among authors in one unified model. The work is motivated by the observation that a link between two documents is not only determined by content similarity, but also affected by the community ties between the authors. Indeed, bloggers are more likely to link to posts in blogs they follow, and researchers are more likely to cite papers presented at the conferences they attend or in the journals they read. This happens because authors are, naturally, more aware of the documents in their community and might not be aware of relevant documents outside it. By accounting for both document similarity and author social network influence on link formation, we can better identify the reasons for the presence or absence of a link, and, in turn, find improved topic models and author communities.

We build our model based on Latent Dirichlet Allocation (LDA), a hierarchical Bayesian model proven

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

successful for topic modeling (Blei et al., 2003). With the additional assumption that the variables representing the community of authors is exchangeable, we can model the membership of authors with a mixture model. Then whether a link exists between two documents follows a binomial distribution parametrized by the similarity between topic mixtures and community mixtures as well as a random factor. We derive inference and estimate parameters using the variational EM approach. In the update equations, we can observe how community information helps to regularize the topic modeling process via citation links and vice versa. To test our model, we apply it to two blog datasets and a subset of the CiteSeer data. We examine both the document topics and community structures our model uncovered and the results demonstrate its effectiveness.

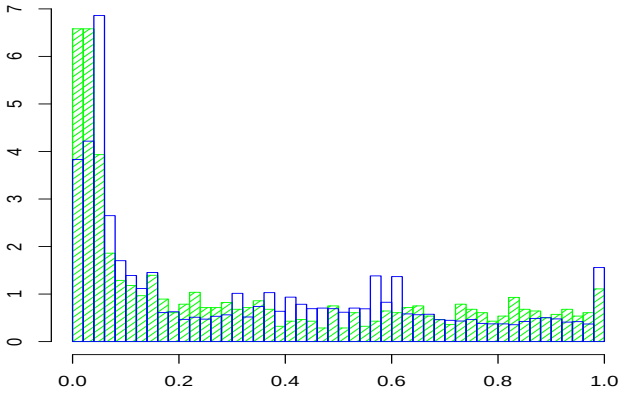

Figure 1. Distributions of content similarity between pairs of documents give the existence of a link between the pair (green shaded bar) or not (blue clear bar). x-axis: similarity score; y-axis: estimated probability density. The similarity score is cosine similarity between topic vectors output by LDA on political blog dataset.

2. Related Work

Topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), have achieved great success in discovering underlying topics from text documents (Figure 2(A)). Recently there has been growing amount of work on developing better topic modeling algorithms with additional information other than texts. One direction is to extend topic models and take into consideration the authorship information. For example, Topic-Author model (Rosen-Zvi et al., 2004) simultaneously models the content of documents and the interest of authors by sharing the hyperparameters of topic mixing for all the documents by the same authors. Later, a Topic-Author-Recipient model (Mccallum et al., 2005) is proposed to consider sender-receiver information so that the distribution over topics is conditioned distinctly on both the sender and the recipient.

Another direction of topic modeling is to explore citation (i.e. link) information. PHITS, an extension to the PLSA model (Cohn & Hofmann, 2001), defines a generative process for both text and citations. It assumes the generation of each hyperlink in a document is a multinomial sampling of the target document from the topic-specific distribution of the documents. A similar model was developed in which PLSA was replaced by LDA as the fundamental generative building block (Erosheva et al., 2004) (Figure 2(B)). Following the convention in (Nallapati & Cohen, 2008), we refer to this model as Link LDA model. Later, Dietz et al. develop the citation influence model to infer the topical influences of citations (Dietz et al., 2007). Nallapati et al. propose Link-PLSA-LDA model as a scalable LDA-type model for topic modeling and link prediction (Nallapati & Cohen, 2008). Very recently, relational topic model (RTM) is introduced to model the link between documents as a binary random variable conditioned on their contents (Chang & Blei, 2009), however, it does not consider community information.

In parallel, graph-based approaches have been extensively studied for community detection (Gibson et al., 1998; Chakrabarti & Faloutsos, 2006) and link prediction (Xu et al., 2005; Yu et al., 2006). As noted earlier, there has been limited work on combining graph and content information for community discovery. In link prediction task, most methods explore both types of information by simply treating the task as a binary classification problem and feeding rich features pre-extracted from texts and link structures as input. To the best of our best knowledge, there is very limited work to jointly model underlying topics, author community, and link formation in one unified model.

3. Topic-Link LDA Model

In this section, we present the Topic-Link LDA model. Before diving into the details of our model, we first investigate the validity of our assumption, i.e. a citation between two documents is not purely due to content similarity. In Figure 1, we show the density of content similarity scores between pairs of blog posts conditioned on whether there is a link between them (see Section 4 for description of the data). The similarity score is calculated as cosine similarity between topic vectors output by LDA. We can observe there are many positive examples with extremely low similarity (close to 0) as well as negative examples with high

similarity scores. To explain away the discrepancy between content similarity and citation, we extend the LDA topic model and combine it with implicit community information. We hope to provide better topic modeling, reveal the community among authors, and supplement current approaches of link prediction.

3.1. Definition of Topic-Link LDA

In the Topic-Link LDA model, we aim to quantify the effect of topic similarity and community similarity to the formation of a link. Therefore the model has three major components with each capturing one perspective of our target. Its graphical model representation is shown in Figure 2(C). As we can see, it has the LDA model for topic modeling on the left, author community model in the middle and link formation model on the right. More specifically, in the author community model, each author is associated with a distribution over community μ, chosen from a Dirichlet (κ). In the topic model, each document is associated with a distribution over topic θ, chosen from a Dirichlet (α). The link formation model associates the existence of a link with a binomial distribution ρ, which is parameterized by the similarity of topic θ and community μ. Notice that in the Author-Topic model and Author-Recipient-Topic model, the Dirichlet parameter α for documents is shared only by documents from the same authors. In our model, we make the simple assumption that all documents share the same Dirichlet parameter so that inference can be made feasible for large-scale datasets. In addition, we argue that if we are only interested in documents within one specific domain (e.g. politics), the simplification might be reasonable.

The next question is how to define a viable similarity measure for topic and community. Several heuristic measures have been discussed in (Dietz et al., 2007), such as Jensen-Shannon-Divergence. Since there are no dominant measures that work well for all text data, we simply adopt dot product as the similarity measure. The final question is how to quantify the effect of topic and community to the formation of a link. In Figure-3, we plot the conditional distribution of a link given topic similarity as a function of topic similarity on the political blog data, i.e. $P(G_{i,j}|\theta_i, \theta_j) = f(\theta_i^T \theta_j)$, where f is either exponential or sigmoid function (these two functions have also been explored in (Chang & Blei, 2009)). It can be seen that both functions are able to capture the main trend of the observations. Therefore we use sigmoid function to demonstrate the derivations of our model (see Table 1 for notations).

To summarize, we have the following data generation

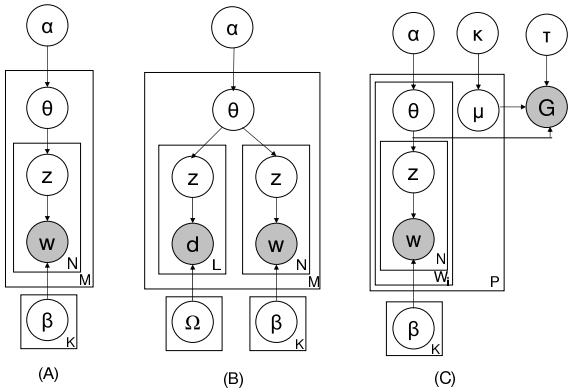

Figure 2. The graphical model representation of (A) Latent Dirichlet Allocation (Blei et al., 2003); (B) Link LDA model (Erosheva et al., 2004); (C) Topic-Link LDA model

process for Topic-Link LDA model

$$\theta|\alpha \sim \text{Dirichlet}(\alpha),\ z|\theta \sim \text{Multi}(\theta),$$
$$w_{i,n}|z, \beta \sim \text{Multi}(\beta), \mu|\kappa \sim \text{Dirichlet}(\kappa),$$
$$G_{i,j}|\mu_{A(i)}, \mu_{A(j)}, \theta_i, \theta_j \sim \text{Bernolli}(\sigma(\rho_{i,j}))$$

where σ is sigmoid function, i.e. $\sigma(x) = 1/(1 + \exp(-x))$, and $\rho_{i,j} = \tau_1 \mu_{A(i)}^T \mu_{A(j)} + \tau_2 \theta_i^T \theta_j + \tau_3$, where A_i is the author index of document i. In other words, we assume the binary variable whether there is a link between two documents follows a binomial distribution with parameter ρ, which is a log-linear function of the content similarity and community similarity.

The joint distribution of observed and hidden variables $P(G, \{w_i\}, \{\theta_i\}, \{z_{i,n}\}, \{\mu_c\}|\alpha, \kappa, \{\beta_k\})$ is:

$$\mathcal{L} = P(G|\{\theta_i\}, \{\mu_c\}) \prod_{i=1}^{P} P(\mu_i|\kappa) \times \quad (1)$$

$$\prod_{i=1}^{M} P(\theta_i|\alpha) \prod_{n=1}^{N} P(z_{i,n}|\theta_{i,n}) \prod_{k=1}^{K} P(w_{i,n}|\beta, z_{i,n,k}) \quad (2)$$

We furthermore define the probability of link graph as

$$P(G|\{\theta_i\}, \{\mu_c\}) = \prod_{i=1}^{M} \prod_{j \neq i}^{M} (\sigma(\rho_{i,j}))^{G_{i,j}} (1 - \sigma(\rho_{i,j}))^{1 - G_{i,j}}$$

As we can see, eq(2) is the complete likelihood of the observed texts and hidden topic variables, and the right-hand side of eq(1) is the complete likelihood of the observed link graph and hidden community variables, which is the major difference between Topic-Link LDA and previous LDA extensions.

3.2. Variational Inference

Similar to LDA, exact inference on the Topic-Link LDA model is intractable. We use the variational

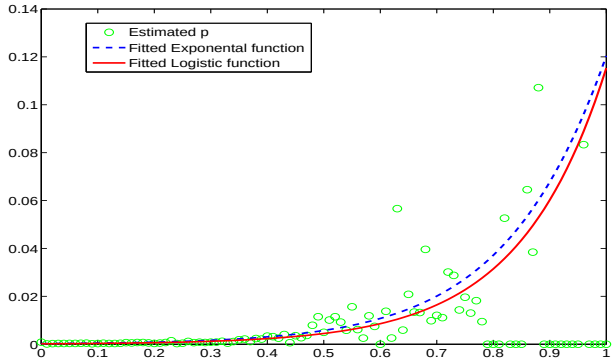

Figure 3. Fitted function for $P(G_{i,j}|\theta_i, \theta_j) = f(\theta_i^T \theta_j)$ (the probability of a link given the similarity). x-axis: similarity score; y-axis: cdf value. Green dots: observations.

method, more specifically the mean-field variational methods, to efficiently obtain an approximation of the objective distribution (Jordan et al., 1999). In short, the mean-field method forms a factorized distribution of the latent variables and fit the parameters of the distribution so that the KL-divergence between the approximate and the true distribution is minimized. The variational distribution we use is:

$$q(\mu, \theta, \mathbf{z}) = q(\mu|\eta)q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n),$$

where $\mu \sim$ Dirichlet(η), $\theta \sim$ Dirichlet(γ), $\mathbf{z_n} \sim$ multinomial(ϕ), and η, γ and ϕ are the free variational parameters. Following the property of Dirichlet distribution, we have $\mathbb{E}_q[\log \mu_h] = \Psi(\eta_h) - \Psi(\sum_{l=1}^{H} \eta_l)$ and $\mathbb{E}_q[\log \theta_h] = \Psi(\gamma_h) - \Psi(\sum_{l=1}^{H} \gamma_l)$, where Ψ is digamma function. We further define $\Delta(\kappa) \equiv \log \Gamma(\sum_{h=1}^{H} \kappa_h) - \sum_{h=1}^{H} \log \Gamma(\kappa_h)$ where Γ is Gamma function.

The expectation of the complete loglikelihood is

$$E_q[\log \mathcal{L}(\alpha, \kappa, \{\beta_k\})] \quad (3)$$

$$= \sum_{i=1}^{M}\sum_{j=i+1}^{M} E_q[(G_{i,j}-1)\rho_{i,j} + \log \sigma(\rho_{i,j})] \quad (4)$$

$$+ \sum_{i=1}^{P}(\Delta(\kappa) + \sum_{h=1}^{H}(\kappa_h - 1)(\Psi(\eta_{i,h}) - \Psi(\sum_{l=1}^{H}\eta_{i,l}))) \quad (5)$$

$$+ \sum_{i=1}^{M}(\Delta(\alpha) + \sum_{k=1}^{K}(\alpha_i - 1)\Psi(\gamma_{i,k}) - \Psi(\sum_{l=1}^{K}\gamma_{i,l}))) \quad (6)$$

$$+ \sum_{i=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} z_{i,n,k}(\Psi(\gamma_{i,k}) - \Psi(\sum_{l=1}^{K}\gamma_{i,l})) \quad (7)$$

$$+ \sum_{i=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{i,n,k}w_{i,n,v}\log\beta_{k,v} \quad (8)$$

The items in eq(5-8) have similar formulation as the LDA model, therefore we focus our discussion on the item in eq(4). As we can see, the challenges involved with computing the term in eq(4) is the expectation of logistic function, i.e. $\sigma(\rho_{i,j})$. To solve the problem, we use the variational methods again with the following variational bound (Jaakkola, 1997):

$$\sigma(x) \geq \sigma(\xi)\exp(\frac{x-\xi}{2} + g(\xi)(x^2 - \xi^2)) \quad (9)$$

where ξ is the free variational parameter and $g(\xi) = (\frac{1}{2} - \sigma(\xi))/2\xi$.

By minimizing the Kullback-Leibler(KL) divergence between the variational distribution q and the true distribution p, we have the following update equations :

$$\phi_{i,n,k} \propto \beta_{iv}\exp(\Psi(\gamma_{i,k}) - \Psi(\sum_{l=1}^{K}\gamma_{i,l})), s.t. \sum_{k}\phi_{i,n,k} = 1;$$

$$\xi_{i,j} = (\mathbb{E}_q[\rho_{i,j}^2])^{\frac{1}{2}}$$

There is no closed form solution for η, and therefore iterative searching algorithms have to be applied where the derivatives can be calculated as Proposition 1 with the constraint that $\sum_h \eta_{i,h} = 1$ and $\eta_{i,h} >= 0$.

Proposition 1: The derivative of loglikelihood with respect to η, $d\eta_{i,j}$ can be calculated as follows:

$$\begin{aligned} d\eta_{i,j} = {} & \Psi'(\eta_{i,h})(\kappa_i - \gamma_{i,h}) - \Psi'(\sum_{l=1}^{H}\eta_{i,l})\sum_{l=1}^{H}(\kappa_i - \gamma_{i,l}) \\ & + (f_1(\eta)^T(Q_1 + Q_1^T))^T : \frac{\partial f_1(\eta)}{\partial \eta_{i,h}} \\ & + (Q_2 + Q_2^T) \circ f_2(\eta)^T : \frac{\partial f_2(\eta)}{\partial \eta_{i,h}} \\ & - (f_3(\eta)^T(Q_2 + Q_2^T))^T : \frac{\partial f_3(\eta)}{\partial \eta_{i,h}} \\ & + (f_4(\eta)^T(Q_2 + Q_2^T))^T : \frac{\partial f_4(\eta)}{\partial \eta_{i,h}} \end{aligned} \quad (10)$$

where

$$\begin{aligned} f_1(\eta)_{i,h} &= \frac{\eta_{i,h}}{\eta_{i,0}}, \qquad f_2(\eta)_{i,j} = \frac{\sum_{h=1}^{H}\eta_{i,h}\eta_{j,h}}{\eta_{i,0}(\eta_{i,0}+1)}, \\ f_3(\eta)_{i,h} &= \frac{\eta_{i,h}\eta_{i,h}}{\eta_{i,0}(\eta_{i,0}+1)}, \; f_4(\eta)_{i,h} = \frac{\eta_{i,h}(\eta_{i,h}+1)}{\eta_{i,0}(\eta_{i,0}+1)}; \\ Q_1(\eta)_{i,j} &= \tau_1(G_{i,j} - \frac{1}{2} + 2g(\xi_{i,j})(\eta_{i,j}^{s_2} + \tau_3)) \\ Q_2(\eta)_{i,j} &= \tau_1 g(\xi_{i,j}) \end{aligned}$$

and $Q_1(\eta)_{i,j} = Q_2(\eta)_{i,j} = 0$ if $i = j$.

The proof and the definition of $\eta_{i,j}^{s_1}$ and $\eta_{i,j}^{s_2}$ can be found in Appendix. We then resort to constrained

Table 1. Notations in the Topic-Link LDA model: Input: Word vector $W_{M\times N} = \{0,1\}$ and Link graph $G_{M\times M} = \{0,1\}$

Parameters	Variational parameters → hidden variables	Constants
$\alpha_{(1\times K)}$	$\gamma \rightarrow \theta_{(M\times K)}$: topic Dirichlet	K: # of hidden topics; M: # of total posts
$\beta_{(K\times N)}$	$\eta \rightarrow \mu_{(P\times H)}$: community Dirichlet	P: # of blogers; H: # of hidden community
$\kappa_{(1\times K)}$	$\xi \rightarrow \rho_{(P\times P)}$: link probability	M_i: # of posts from the i_{th} blogger
$\tau_{(1\times 3)}$	$\phi \rightarrow Z_{(M\times K\times N)}$: topic indictors	N: # of words in the vocabulary

optimization (active set) algorithm and line search to compute the value of η. γ has similar form as η and details are omitted.

In previous discussion, we motivate the advantage of Topic-Link LDA model as one unified framework that jointly model texts and author community. From the updating equation (10), we can easily observe how the topic information serve as a regularizer when the model infers the hidden communities and vice versa.

3.3. Parameter Estimation

To estimate the parameters of our model, we use the variational EM procedure that maximizes a lower bound provided by the variational method with respect to the parameters α, κ, β and τ. Taking the derivatives of the lower bound and setting it to zero, we have the following update equations for β and τ:

$$\beta_{k,v} \propto \sum_{i=1}^{M}\sum_{n=1}^{N}\phi_{i,n,k}w_{i,n,v},$$

$$\tau_1 = -\frac{\sum_{i=1}^{P}\sum_{j\neq i}^{P}(G_{i,j} - \frac{1}{2} + 2g(\xi_{i,j})(\tau_2\gamma_{i,j}^{s_1} + \tau_3))\eta_{i,j}^{s_1}}{2\sum_{i=1}^{P}\sum_{j\neq i}^{P} g(\xi_{i,j})\eta_{i,j}^{s_2}}$$

τ_2 and τ_3 has similar forms as τ_1 and therefore we omit the details. There is no closed form solution for α. We applied iterative searching algorithms, where the derivatives are as follows:

$$\frac{\partial \log \mathcal{L}}{\partial \alpha_k} = M(\Psi(\sum_{l=1}^{K}\alpha_l) - \Psi(\alpha_k)) + \sum_{i=1}^{M}(\Psi(\gamma_{i,k}) - \Psi(\sum_{l=1}^{K}\gamma_{i,l}))),$$

$$\frac{\partial \log \mathcal{L}}{\partial \alpha_k \alpha_t} = M\Psi'(\sum_{l=1}^{K}\alpha_l) - \delta(k,t)M\Psi'(\alpha_k).$$

κ has the same derivation as α and we omit the details.

3.4. Link Prediction

One natural result of the Topic-Link LDA model is to predict the probability of a link between two documents given the previous documents and their link structure. For example, in blog analysis, we might be interested in predicting whether a new post will be developed into a well-cited post (measured by the number of citation links) in the future based on past blog posts and their link structures; in movie recommendation, we might be interested in answering how many reviews a new movie will get based on previous rating history as well as text description of movies. To solve the problem, we need to make two assumptions about the data, including (1) the author (or user) community remains the same for training and testing data; (2) how content similarity and community similarity contribute to the formation of a link is the same for training and testing. We argue that these two assumptions are reasonable and very common in current algorithms for link prediction.

The link prediction algorithm works as follows: we run Topic-Link LDA over training data and compute the values of community parameter μ and κ as well as link formation parameter τ. For a pair of testing documents (i, j), we compute the topic modeling parameter θ, calculate the probability of a link as $\rho = \sigma(\tau_1\mu_{A(i)}^T\mu_{A(j)} + \tau_2\theta_i^T\theta_j + \tau_3)$, and predict that a link exists if ρ is above some threshold.

4. Experiment Results

To examine the effectiveness of our Topic-Link LDA model, we use (1) two blog data sets from different subject domains, including web 2.0 technologies and US politics respectively, and (2) research paper citation data from CiteSeer. For the blog data, we select a list of around 100 blogs who are most influential in each domain based on external forms of measure (Notice that no information about the blog communities was used in constructing the data set). These blogs are scanned daily and we use the blog posts within Feb 1-14, 2008 as the training set and those within Feb 15-22, 2008 as test set. **Web 2.0 Blogs** dataset consists of 3853 and 2096 posts for training and testing respectively, focusing on web 2.0, internet technologies, and technology start-ups. We select the top 75 blogs as listed by Technorati and Techmeme Leaderboard. It includes popular sites like BoingBoing, Engadget, Lifehacker, and TechCrunch. **Political Blogs** dataset

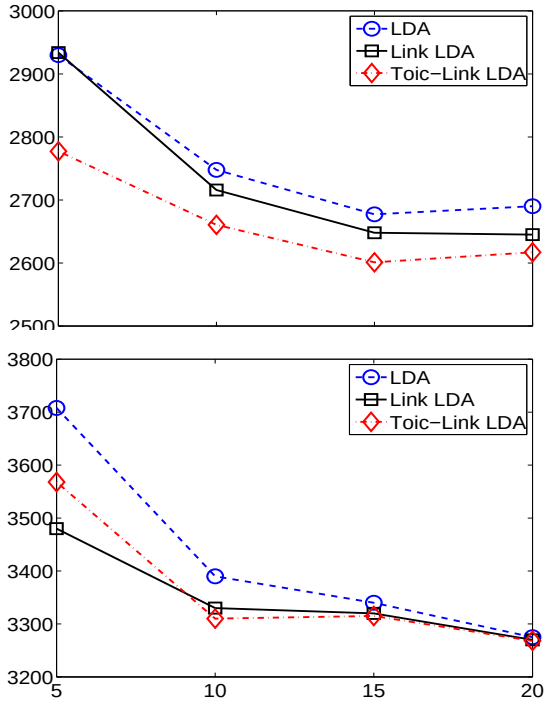

Figure 4. The perplexity of LDA styled model on Web2.0 blog (Top) and political blog (Down). X-axis: number of hidden topics; Y-axis: perplexity on the test set

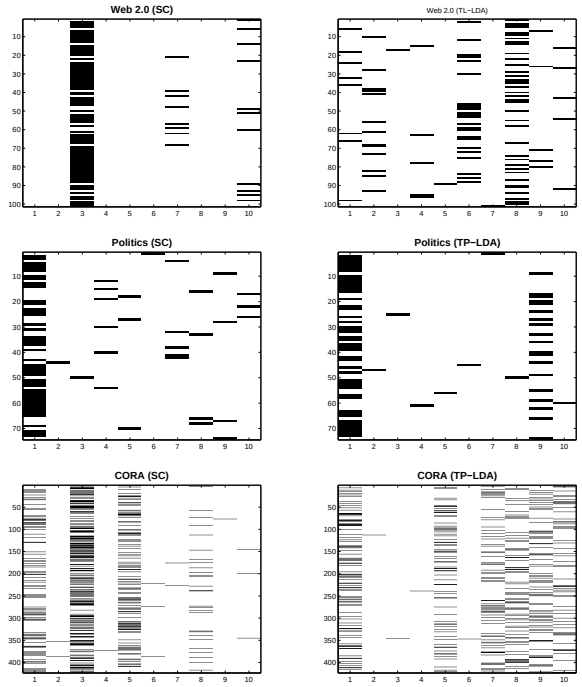

Figure 5. Discovered author community by spectral clustering and Topic-Link LDA

consist of 3467 and 1897 posts for training and testing respectively. Similar as Web 2.0 blogs, 101 political blogs are chosen based on the rankings at Technorati, as well as most popular blog listings in literature. For paper citation data, we use a subset of the **CORA** data set (McCallum et al., 2000). The abstract and title of a paper is treated as its content and the reference information as link. We choose authors who published papers classified as "Artificial Intelligence-Machine Learning" and "Artificial Intelligence-Data Mining" as our analysis focus, which results in 423 authors with their 2695 papers. We pruned the vocabulary by stemming each term to its root, removing stop words, and also removing terms that occurred fewer than 2 times, resulting in a vocabulary of around 10,000 for the two blog datasets and 3,000 for CORA dataset.

Topic Modeling To demonstrate the effectiveness of Topic-Link LDA model on topic modeling, we compute the perplexity of the test sets in Web 2.0 Blog and Political Blog data, with parameters learned from the corresponding training sets. To see if modeling the author community structure provides any benefit we also train a Topic-Link LDA model where we restrict the number of communities to be one. Figure-4 shows the perplexity scores on different number of hidden topics for LDA, Link LDA (a single author community) and Topic-Link LDA with ten author communities. We can see that on Web 2.0 data, Topic-Link LDA yields lower perplexity than both the LDA and Link LDA models. On the Political Blog data the perplexity results for the three types of models are comparable. As we will see in the next section, however, Topic-Link LDA uncovers a very interesting community structure for this data set. In addition, we show the top 5 topics (with the largest number of associated blog posts) among the 15 topics learned from the LDA and Topic-Link LDA on CORA dataset in Table 2. The topics from the Topic-Link LDA model are very reasonable, which include major areas in the community, i.e. reinforcement learning, Bayesian network, classification and genetic algorithm, as well as methodology description topic (i.e. topic 2). In contrast, LDA splits reinforcement learning and Markov model into two separate topics while our algorithm identifies them as one topic by making use of the citation information.

Community Discovery One advantage of the Topic-Link LDA model is its ability to uncover the author community with information from both link and content. Figure 5 shows the grouping results of blogger or authors in our data using spectral clustering with citation graph and our model. The number of clusters is set to 10 for both methods. We can see that

Table 2. An illustration of the top 5 topics from the 15 topics for CORA dataset. Each topic is shown with 7 words with the highest conditional probability.

	Topic Rank	LDA	Topic-Link LDA
CORA	1	reinforc, use, algorithm, optim, learn, process	markov, model, stat, algorithm, reinforc, support
	2	model, report, markov, bayes, statist, chain, sampl	report, univer, infer, prob, network, scient, bayes
	3	algorithm, learn, bound, class, numb, time, set	learn, pape, research, present, task, machin, discuss
	4	genet, evolut, algorithm, problem, use, program, behavior	learn, method, algorithm, classif, train, problem, result
	5	network, neur, learn, connect, input, repret, model, featur	genet, evolut, algorithm, pape, result, search

Table 3. Examples of identified communities from CORA data set by Topic-Link LDA

Group 1	Group 2	Group 3	Group 4
E Sontag	S Singh	T Sejnowski	J Shavlik
N Friedman	D MacKay	M Jordan	J Pearl
H Tirri	N Intrator	R Sutton	V Honavar
C Boutilier	J Schmidhuber	G Hinton	C Lee
Y Freund	A Weigend	G Cottrell	T Dietterich
P Myllymaki	M Wellman	Z Ghahramani	R Kohavi
R Tibshirani	R Neal	M Mozer	S Thrun
S Lawrence	A Raftery	O Mangasarian	P Chan
S Salzberg	W Buntine	B Krose	M Kearns
Y Singer	F Girosi	J Marshall	W Cohen

Topic-Link LDA model tends to have more balanced clusters thanks to the Dirichlet prior. In addition, we argue its results are more meaningful. For example, for political blogs, the citation graph is sparse and the connection is mostly chain structure, therefore spectral clustering outputs one big cluster with the majority of the bloggers. In contrast, our algorithm splits the bloggers into two large clusters, which seem to be in line with the political affinities of the bloggers (i.e. democrats vs. republicans). Table 3 shows some examples of communities that our algorithm identified from the CORA citation data.

It is interesting to examine how content similarity and community similarity contribute to the formation of a link. We define the contribution of community similarity (or content similarity) as the product of its coefficient τ_1 (or τ_2 for content similarity) and the mean of corresponding similarity scores of all training examples. In Figure 6 we plot the ratio between the contribution of community similarity and content similarity for the three domains. We can observe that all values are less than 1, which indicates that content similarity generally plays a more important role in link formation. Furthermore, it is interesting to observe that author community has much stronger effect to link formation in political domains than technical domain and scientific papers.

Link Prediction As discussed in Section 3.4, we can also use the results from the Topic-Link LDA model for link prediction. We test the algorithm on Web 2.0 and Political Blog data and compare it to *graph-based* preferential attachment method and *content-based* supervised learning method with content similarity between input pairs of posts as features and logistic regression as classifiers. Results in terms of precision, recall and F1 are presented in Table 4. We can see that the link prediction task is extremely challenging for these data sets. While Topic-Link LDA is able to yield more reasonable results than the two competing methods, the precision is still low for practical purposes.

Table 4. Results of link prediction (Precision, Recall and F1) on two blog datasets

	Topic-Link LDA			Graph-based			Content-based		
Web 2.0	.056	.724	.103	.001	.836	.002	.020	.533	.030
Politics	.079	.803	.144	.005	.891	.010	.020	.750	.039

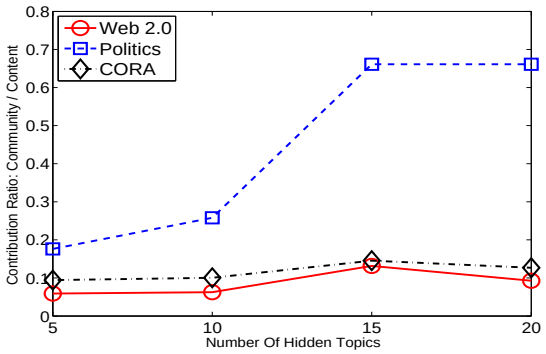

Figure 6. The ratio between the contribution of community similarity and that of content similarity across datasets

5. Conclusion

In this paper, we develop the Topic-Link LDA model to jointly model topics and author community. It assumes the formation of a link between two documents as a combination of topic similarity and community closeness, which brings both topic modeling and community discovering in one unified model. The experiment results have demonstrated the effectiveness of our algorithm. For future work, we are interested in extending the model to analyze time-series linked documents. In addition, since some authors are naturally more influential in their community, it would be interesting to consider the dynamics between citation and influence as topics evolve over time.

Acknowledgments

We thank John Lafferty, Jure Leskovec, Ramesh Nallapati for discussing the ideas in the paper. We thank anonymous reviewers for their valuable suggestions.

References

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proc. of Int. Conf. on Mach. Learn. (ICML'06)* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Chakrabarti, D., & Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, *38*, 2.

Chang, J., & Blei, D. (2009). Relational topic models for document networks. *Proc. of Conf. on AI and Statistics (AISTATS'09)*.

Cohn, D., & Hofmann, T. (2001). The missing link - a probabilistic model of document content and hypertext connectivity. *Proc. of Conf. on Neural Information Processing Systems (NIPS'01)* (pp. 430–436).

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. *Proc. of Int. Conf. on Mach. Learn. (ICML'07)* (pp. 233–240).

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed membership models of scientific publications. *Proc. Nat. Acad. Sci.*, *101*, 5220–5227.

Gibson, D., Kleinberg, J. M., & Raghavan, P. (1998). Inferring web communities from link topology. *UK Conference on Hypertext* (pp. 225–234).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proc. Nat. Acad. Sci.*, *101*, 5228–5235.

Jaakkola, T. (1997). Variational methods for inference and estimation in graphical models. *PhD thesis, MIT*.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233.

Mccallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *Proc. of Int. Joint Conf. on Articial Intelligence (IJCAI'05)* (pp. 786–791).

McCallum, A., Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, *3*, 127–163.

Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. *Proc. of Int. World Wide Web Conf. (WWW'08)* (pp. 101–110).

Nallapati, R., & Cohen, W. (2008). Link-plsa-lda: A new unsupervised model for topics and influence in blogs. *Proc. of Int. Conf. on Weblogs and Social Media (ICWSM'08)* (pp. 84–92).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proc. of Conf. on Uncertainty in Artificial Intelligence (UAI'04)* (pp. 487–494).

Xu, Z., Tresp, V., Yu, K., Yu, S., & Kriegel, H.-P. (2005). Dirichlet enhanced relational learning. *Proc. of Int. Conf. on Mach. Learn.* (pp. 1004–1011).

Yu, K., Chu, W., Yu, S., Tresp, V., & Xu, Z. (2006). Stochastic relational models for discriminative link prediction. *Proc. of Conf. on Neural Information Processing Systems (NIPS'06)* (pp. 1553–1560).

Appendix: Proof of Proposition 1

Applying the bound in eq(9)to eq(4), we have

$$\mathbb{E}_q[\log \mathcal{L}(4)] \geq \sum_{i=1}^{M}\sum_{j=i+1}^{M} \{(G_{i,j}-1)\mathbb{E}_q[\rho_{i,j}] \log\sigma(\xi_{i,j}) + \frac{\mathbb{E}_q[\rho_{i,j}]-\xi_{i,j}}{2} + g(\xi_{i,j})(\mathbb{E}_q[\rho_{i,j}^2]-\xi_{i,j}^2)\}.$$

By definition, $\rho_{i,j} = \tau_1 \mu_{A(i)}^T \mu_{A(j)} + \tau_2 \theta_i^T \theta_j + \tau_3$, then

$$\begin{aligned}
\mathbb{E}_q[\rho_{i,j}] &= \tau_1 \eta_{i,j}^{s_1} + \tau_2 \gamma_{i,j}^{s_1} + \tau_3 \\
\mathbb{E}_q[\rho_{i,j}^2] &= \tau_1^2 \,\mathrm{Tr}[\eta_{i,j}^{s_2}] + \tau_2^2 \,\mathrm{Tr}[\gamma_{i,j}^{s_2}] \\
&\quad + 2\tau_1\tau_2 \eta_{i,j}^{s_1}\gamma_{i,j}^{s_1} + 2\tau_1\tau_3\eta_{i,j}^{s_1} + 2\tau_2\tau_3\gamma_{i,j}^{s_1} + \tau_3^2
\end{aligned}$$

where $\eta_{i,j}^{s_1} = \mathbb{E}_q[\eta_{i,:}]\mathbb{E}_q[\eta_{j,:}]^T$, $\eta_{i,j}^{s_2} = \mathbb{E}_q[\eta_{i,:}\eta_{i,:}]\mathbb{E}_q[\eta_{j,:}\eta_{j,:}]^T$. Following the property of Dirichlet distribution, we have

$$\begin{aligned}
\mathbb{E}_q[\eta_{i,h}] &= \frac{\eta_{i,h}}{\eta_{i,0}}, \\
\mathbb{E}_q[\eta_{i,h}\eta_{i,l}] &= \begin{cases} \frac{\eta_{i,h}\eta_{i,l}}{\eta_{i,0}(\eta_{i,0}+1)} & \text{if } h \neq l, \\ \frac{\eta_{i,h}(\eta_{i,h}+1)}{\eta_{i,0}(\eta_{i,0}+1)} & \text{if } h = l. \end{cases}
\end{aligned}$$

where $\eta_0 = \sum_k \eta_k$. The derivation also holds for γ. The terms associated with $\eta_{i,h}$ in log-likelihood are:

$$\begin{aligned}
&\log \mathcal{L}_{[\eta_{i,h}]} = \\
&\sum_{i=1}^{P}\sum_{h=1}^{H}\{(\Psi(\eta_{i,h}) - \Psi(\sum_{l=1}^{H}\eta_{i,l}))(\kappa_i - \eta_{i,h}) - \Delta(\eta)\} \\
&+\mathrm{Tr}[f_1(\eta)^T Q_1 f_1(\eta) + f_2(\eta) \circ Q_2 * f_2(\eta)] \\
&-\mathrm{Tr}[f_3(\eta)^T * Q_2 * f_3(\eta) + f_4(\eta)^T Q_2 f_4(\eta)]. \quad (11)
\end{aligned}$$

Taking the derivative of eq(11), we get Proposition 1.