Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks

Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, Cindy Xide Lin
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{hbdeng, hanj, bozhao3, yintao, xidelin2}@uiuc.edu

ABSTRACT

With the development of Web applications, textual documents are not only getting richer, but also ubiquitously interconnected with users and other objects in various ways, which brings about text-rich heterogeneous information networks. Topic models have been proposed and shown to be useful for document analysis, and the interactions among multi-typed objects play a key role at disclosing the rich semantics of the network. However, most of topic models only consider the textual information while ignore the network structures or can merely integrate with homogeneous networks. None of them can handle heterogeneous information network well. In this paper, we propose a novel topic model with biased propagation (TMBP) algorithm to directly incorporate heterogeneous information network with topic modeling in a unified way. The underlying intuition is that multi-typed objects should be treated differently along with their inherent textual information and the rich semantics of the heterogeneous information network. A simple and unbiased topic propagation across such a heterogeneous network does not make much sense. Consequently, we investigate and develop two biased propagation frameworks, the biased random walk framework and the biased regularization framework, for the TMBP algorithm from different perspectives, which can discover latent topics and identify clusters of multi-typed objects simultaneously. We extensively evaluate the proposed approach and compare to the state-of-the-art techniques on several datasets. Experimental results demonstrate that the improvement in our proposed approach is consistent and promising.

Categories and Subject Descriptors:H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; H.2.8 [**Information Systems Applications**]: Database Applications—*Data mining*

General Terms: Algorithm, Experimentation

Keywords: Topic modeling, biased propagation, clustering, heterogeneous information network

KDD'11, August 21–24, 2011, San Diego, California, USA.

1. INTRODUCTION

In this paper, we study the problem of topic modeling and object clustering on text-rich *heterogeneous information networks*. Textual documents, such as web pages, papers and blogs, are ubiquitously interconnected with each other as well as with other objects (e.g., users) in various ways, leading to simultaneous growth of both textual documents and heterogeneous network structures between documents and other objects. Information networks have been popularly used to represent networked systems, and a text-rich heterogeneous information network is formed when the network consists of a large number of text data as well as other objects. Taking bibliographic data as an example, as researchers are regularly publishing papers in various venues (e.g., conferences, journals, etc.), we not only obtain textual information of documents, but also have access to the intersections among multi-typed objects such as documents, authors and venues as shown in Figure 1(a). Figure 1(b) illustrates a simplified and basic heterogeneous information network with two types of objects: the documents D with rich text and the associated users U without explicit text, for example, blogs and bloggers, webpages and online users. There are many other text-rich information network examples that consist of a large number of interacting, multi-typed components accompanying with rich text data.

These examples show that in reality we are dealing with collections of documents as well as other objects in a heterogeneous information network. Therefore, it is important and challenging to examine how text data and heterogeneous information network can mutually enhance each other in topic modeling and other text mining tasks. With multi-typed objects accompanying with text documents, it is highly desirable to analyze how topics propagate from documents to other objects, and how the topics of other objects reinforce topic modeling and object clustering simultaneously.

Many topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [13] and Latent Dirichlet Allocation (LDA) [3], have been proposed and shown to be useful for document analysis, but most of them only consider the textual information while ignore the network structures. Recently, several studies, including NetPLSA [17], Laplacian PLSI [6] and Locally-consistent Topic Model [7], have been proposed for combining topic modeling and network structures. However, these models can merely deal with homogeneous networks, such as document nearest-neighbor graph and co-authorship graph, but not heterogeneous information networks. Although a heterogeneous information network can be transformed into or be regarded as a homogeneous

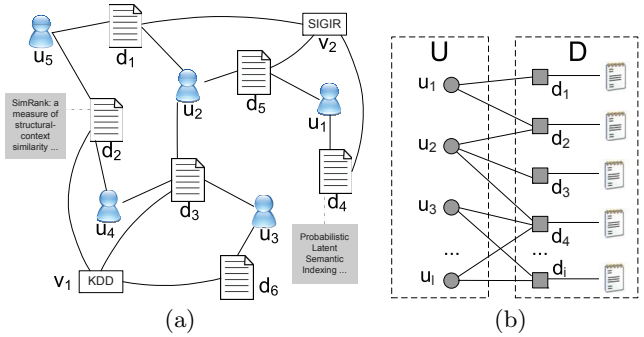


Figure 1: Examples of heterogeneous information networks.

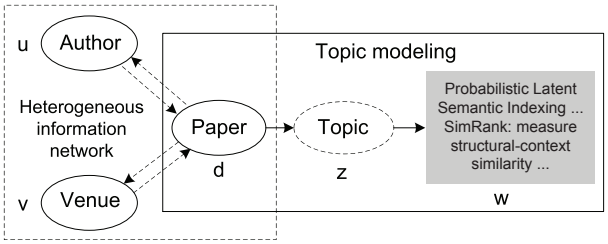


Figure 2: Topic modeling with heterogeneous information network.

information network, this may result in loss of rich semantics of the original network. Moreover, different objects have their own inherent information, which should be treated differently. Therefore, it is reasonable and challenging to directly incorporate the heterogeneous information network with topic modeling in a unified framework.

To address the problem, we propose a novel Topic Model with Biased Propagation algorithm (TMBP in short) which can be used to discover latent semantic topics and reinforce clusters of multi-typed objects simultaneously. Consequently, we investigate two alternative frameworks, i.e., biased random walk framework (TMBP-RW) and biased regularization framework (TMBP-Regu), for incorporating with topic modeling from different views. The basic idea of the biased random walk framework is to propagate the topic probabilities obtained by topic models on the heterogeneous information network via a biased propagation, as illustrated in Figure 2. A simple and unbiased topic propagation across different objects on such a heterogeneous network does not make sense. The underlying intuition is that different objects should be treated differently along with their inherent information. For example, as shown in Figure 1(b), the topic of a document d_i can be identified by mining its text information, while the interest of a user u_l without explicit text information can be characterized simply based on the associated documents which is captured by the heterogeneous network. On the other hand, the estimated interest of a user may affect the topic of a document afterward. In this way, there is a naturally biased topic propagation and consistency across different objects. Furthermore, we develop a joint regularization framework to incorporate a heterogeneous network into topic modeling by regularizing a statistical topic model along with a biased regularization on the heterogeneous information network. The biased regularization framework exploits valuable and reinforced information from heterogeneous network and provides promising constraints of overfitting for topic modeling, which leads to a significant improvement over the baseline method. Finally, we conduct extensive experiments and compare with the state-of-the-art techniques for object clustering and topic modeling tasks using two real-world datasets. Experimental results show that our TMBP-Regu model achieves the best performance.

In a nutshell, our contributions of this paper are: (1) the introduction of the TMBP algorithm to directly incorporate heterogeneous information network instead of homogeneous information network with topic modeling; (2) the investigation of two biased propagation frameworks, including the biased random walk framework and the biased regularization framework, for the TMBP algorithm from different perspectives; (3) the biased regularization term for treating documents with rich text and other objects without explicit text in a different way; and (4) the application of our model for clustering multi-typed objects collaboratively, in which TMBP-Regu achieves the best performance.

The rest of this paper is organized as follows. We first introduce the preliminaries in Section 2. Section 3 briefly describes probabilistic topic models. In Section 4, we systematically present and develop the proposed TMBP algorithm. Extensive experimental results on object clustering and document modeling are reported in Section 5. Finally, we review some related work in Section 6, and present our conclusions and future work in Section 7.

2. PRELIMINARIES

In this section, we formally introduce several related concepts and notations, and define the problem of topic modeling and object clustering in a text-rich heterogeneous information network. We assume that the data to be analyzed consists of both a collection of text documents and an associated heterogeneous network with multi-typed objects.

Definition 1 (Information Network): An information network consists of T types of objects $\mathcal{X} = \{X_t\}_{t=1}^{T}$, where X_t is a set of objects belonging to t_{th} type. Such a network with different types of objects can be denoted as a graph $G = (\mathcal{X}, E)$, where $\mathcal{X}$ is a set of vertices representing objects, i.e., $\mathcal{X} = X_1 \cup X_2 \cup ... \cup X_T$, and E is a set of edges representing the relation between objects. Suppose x is a vertex $x \in \mathcal{X}$, an edge $\langle x_i, x_j \rangle$ is a binary relation between vertices x_i and x_j. Specially, the network is called **homogeneous information network** when $T = 1$; and it becomes **heterogeneous information network** when $T \geq 2$.

A **text-rich heterogeneous information network** is formed when the information network contains a set of text documents $D = \{d_1, d_2, ..., d_N\}$ and several other types of objects, which is denoted as $\mathcal{X} = D \cup \{X_t\}_{t=1}^{T-1}$. Each document is represented as a bag of words, i.e., $d = \{w_1, w_2, ..., w_{|d|}\}$, and we use $n(d_i, w_j)$ to denote the occurrences of word w_j in d_i. For the bibliographic network as shown in Figure 1(a), there are three types of objects (i.e., $T = 3$), including papers D, authors U and venues V. The text-rich information network can be denoted as $G = (D \cup U \cup V, E)$, where E is the set of edges representing the relationships between documents D and objects U, V. For simplicity, in this paper we mainly consider $T = 3$, but the proposed model can be easily extended to incorporate more types of objects.

Now we can formulate our *topic modeling and object clustering* problem as: Given a document collection D and a text-rich information network $G = (D \cup U \cup V, E)$, the task

of topic modeling is to model and extract K major topic models $Z = \{z_1, z_2, ..., z_K\}$ associated with multi-typed objects, where K is a user specified parameter. A latent topic model z_k is a probabilistic distribution of words in the vocabulary of collection. The probability of a word w is referred as $P(w|z)$. The task of object clustering is to group different types of objects into proper clusters simultaneously.

3. PROBABILISTIC TOPIC MODELS

Topic modeling has been popularly used for data analysis in various domains [3, 13, 17, 22]. A number of recent approaches [3, 13] to modeling document content are based upon the idea that the probability distribution over words in a document can be expressed as a mixture model of K topics, where each topic is a probability distribution over words. We will describe one of the most well-known and fundamental topic models, Probabilistic Latent Semantic Analysis (PLSA) [13]. In PLSA, an unobserved topic variable $z_k \in \{z_1, ..., z_K\}$ is associated with the occurrence of a word $w_j \in \{w_1, ..., w_M\}$ in a particular document $d_i \in \{d_1, ..., d_N\}$. By summing out the latent variable z, the joint probability of an observed pair (d, w) can be defined as

$$P(d_i, w_j) = P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i), \quad (1)$$

where $P(w_j|z_k)$ is the probability of word w_j according to the topic model z_k, and $P(z_k|d_i)$ is the probability of topic z_k for document d_i. Following the likelihood principle, these parameters can be determined by maximizing the log likelihood of a collection C as follows:

$$\mathcal{L}(C) = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i). \quad (2)$$

The model parameters $\phi = \{P(w_j|z_k)\}$ and $\theta = \{P(z_k|d_i)\}$ can be estimated by using standard EM algorithm [9].

Thus PLSA provides a simplified solution to model topics of documents in a text-rich information network. However, this model ignores the associated heterogeneous information network as well as other interacted objects, so it cannot model and make use of associated objects simultaneously. Another limitation of PLSA is that there is no constraint on the parameters $\theta = \{P(z_k|d_i)\}$, the number of which grows linearly with the data. Therefore, the model is prone to overfitting the data. To alleviate these problems, we propose the following biased topic propagation algorithm by exploiting the heterogeneous information network.

4. BIASED TOPIC PROPAGATION

In this section, we propose a novel and general biased topic propagation algorithm to incorporate the heterogeneous information network with the textual information for topic modeling, so as to estimate the probabilities of topics for documents as well as other associated objects and improve the performance of topic modeling simultaneously.

4.1 Biased Random Walk Framework

In order to obtain the topics for other objects, a straightforward way is to propagate the topic probabilities from documents to other objects through the heterogeneous information network as shown in the dashed rectangle of Figure 2. The basic criterion is that different objects which are connected with each other should have similar weights of topics. To be more specific, the topic distribution of an object without explicit text information (e.g., u_l in Figure 1(b)) depends on the topic distribution of the documents it connects. For example, the research topic of an author could be characterized by his/her published papers, and the interest of a blogger is highly correlated with the associated blog posts. On the other hand, the topic of a document is also correlated with its authors to some extent, but, most importantly, its topic should be principally determined by its inherent content of the text. Therefore, different objects and interactions reflect distinctive semantics of a heterogeneous network, which should be treated differently. So we propose a biased random walk framework to cope with such a heterogeneous information network.

Let us first take Figure 1(b) as an example and discuss how the topics propagate from documents to neighboring objects. Given the topic probabilities of documents $P(z_k|d_i)$, the probabilities for a user u can be calculated by:

$$P(z_k|u) = \sum_{d_i \in \mathcal{D}_u} P(z_k|d_i)P(d_i|u) = \sum_{d_i \in \mathcal{D}_u} \frac{P(z_k|d_i)}{|\mathcal{D}_u|}, \quad (3)$$

where $\mathcal{D}_u$ is a set of documents that are associated with user u, and $|\mathcal{D}_u|$ is the number of documents (i.e., the degree of user u in the graph). The underlying intuition behind the above equation is that the topic distribution of an object is determined by the average topic distribution of connected documents. Similarly, the probabilities for a venue $v \in \mathcal{V}$ in Figure 1(a) can be defined as:

$$P(z_k|v) = \sum_{d_i \in \mathcal{D}_v} P(z_k|d_i)P(d_i|v) = \sum_{d_i \in \mathcal{D}_v} \frac{P(z_k|d_i)}{|\mathcal{D}_v|}, \quad (4)$$

where $\mathcal{D}_v$ is a set of documents that are published in venue v. Since many objects usually do not have explicit text content for further representation, e.g., users and venues in Figure 1(a), their topic distributions are entirely dependent on the estimated topic distributions of connected documents.

On the other hand, the topic distributions could be propagated from these objects to documents, so as to reinforce the topic distributions of documents. Along with the inherent topic probabilities estimated from the text, we propose the following biased topic propagation

$$P(z_k|d) = \xi P(z_k|d) + \frac{(1-\xi)}{2} \left(\sum_{u \in \mathcal{U}_d} \frac{P(z_k|u)}{|\mathcal{U}_d|} + \sum_{v \in \mathcal{V}_d} \frac{P(z_k|v)}{|\mathcal{V}_d|} \right) \quad (5)$$

where $\mathcal{U}_d$ represents the set of authors of paper d, and $\mathcal{V}_d$ is the venue associated with paper d. Note that ξ is the biased parameter to control the balance between inherent topic distribution $P(z_k|d)$ and the propagated topic distribution. If $\xi = 1$, the topics of documents retain the original ones, while the topics of other objects are determined by their associated documents in one step. We refer this special case as our **baseline** model with PLSA.

According to Eqs. (3), (4) and (5), we formulate the biased random walk framework at the topic level. The final topic probabilities of different objects can be obtained through an iteratively updated process. Here we focus on topic propagation between different objects on heterogeneous networks, but the propagation on homogeneous networks, such as citation graph, can be easily integrated into this framework.

4.2 Biased Regularization Framework

In the previous section, we give a way to biased topic propagation algorithm, but the topic modeling and random walk process are combined as two independent stages, so they can not mutually enhance each other. Here we investigate a joint regularization framework to directly incorporate heterogeneous information network into topic modeling by regularizing a statistical topic model with a biased regularizer on the heterogeneous information network.

In this section, we take bibliographic information network as a concrete example illustrated in Figure 2 to show how this idea works. Let us first discuss the paper-author bipartite graph. Generally, an author $u_l \in \mathcal{U}$ is knowledgeable about a specific topic if author u_l has published papers related to the topic. Similarly, a paper $d_i \in \mathcal{D}$ may be related to a specific topic if its authors have knowledge in that area. Thus, we define a regularization term as:

$$\begin{aligned} R_U &= \frac{1}{2}\sum_{i=1}^{|\mathcal{D}|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_l \in \mathcal{U}_{d_i}} \frac{P(z_k|u_l)}{|U_{d_i}|}\right)^2 \\ &+ \frac{\tau}{2}\sum_{l=1}^{|\mathcal{U}|}\sum_{k=1}^{K}\left(P(z_k|u_l) - \sum_{d_i \in \mathcal{D}_{u_l}} \frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|}\right)^2. \end{aligned}$$

A natural explanation of minimizing R_U is that authors should have similar topic distribution with their papers, and vice versa. Note that τ is the biased parameter. When $\tau \to \infty$, minimizing R_U will ensure the hypothesis that objects without explicit textual information are completely dependent on the estimated topic distributions of connected documents. Then the objective function R_U can be rewritten as

$$R_U = \frac{1}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_l \in U_{d_i}} \frac{P(z_k|u_l)}{|U_{d_i}|}\right)^2 \quad (6)$$

$$s.t. \quad P(z_k|u_l) - \sum_{d_i \in \mathcal{D}_{u_l}} \frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|} = 0. \quad (7)$$

Thus we formulate the biased regularization term for treating documents with rich text and other objects without explicit text in a different way.

Similarly, the research topic of a venue could be represented by the published papers of the venue, and the topic of venues will guide relevant papers to be submitted or published in the corresponding venue. Thus, we could define the regularization term between documents and venues as:

$$R_V = \frac{1}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{v_m \in \mathcal{V}_{d_i}} \frac{P(z_k|v_m)}{|\mathcal{V}_{d_i}|}\right)^2 \quad (8)$$

$$s.t. \quad P(z_k|v_m) - \sum_{d_i \in \mathcal{D}_{v_m}} \frac{P(z_k|d_i)}{|\mathcal{D}_{v_m}|} = 0. \quad (9)$$

Minimizing function R_V will smooth the topic distributions between documents and their associated venues, making them more similar.

When a text-rich information network involves several different interactions between multi-typed objects, a joint regularization term $R(G)$ can be defined to combine all these regularization terms together as $R(G) = R_U + R_V$. Intuitively, $R(G)$ measures the difference of the topic models between documents and other objects for each explicit relationship embedded in a heterogeneous network. The more they differ, the larger $R(G)$ would be. So it can be regarded as a "loss function" to help us assess how well the topic distributions on the heterogeneous graph are consistent and correlated semantically. Clearly, we would like the extracted topics to have a small $R(G)$. Actually, the joint regularization term $R(G)$ is very general, which can be straightforwardly extended to consider other graph information.

To incorporate both the textual information and the heterogeneous network, we define a biased regularization framework by adding the regularization term to the log-likelihood

$$\begin{aligned} \mathcal{L} &= L(C) - \lambda R(G), \quad (10) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \\ &\quad -\frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_j \in \mathcal{U}_{d_i}} \frac{P(z_k|u_j)}{|\mathcal{U}_{d_i}|}\right)^2 \\ &\quad -\frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{v_j \in \mathcal{V}_{d_i}} \frac{P(z_k|v_j)}{|\mathcal{V}_{d_i}|}\right)^2, \end{aligned}$$

along with the constraints as defined in Eqs. (7) and (9). In Eq. (10), $L(C)$ measures how likely the data is generated from the topic model based on the collection of documents, and λ is the regularization parameter which is used to control the balance between the data likelihood and the smoothness of topic distributions over the heterogeneous network. It is easy to show that if $\lambda = 0$, the regularized topic model boils down to the standard PLSA. If $\lambda > 0$, the biased regularization model takes into account both the textual information and the heterogeneous relationships across multi-typed objects, which will provide valuable constraints of overfitting for PLSA so as to improve the performance of topic modeling. In the following section, we discuss parameter estimation of the biased regularization framework.

4.3 Model Fitting with Generalized EM

Our parameters include all the topics and the distributions of topics in all objects including documents, authors and venues, which we denote by $\phi = \{P(w_j|z_k)\}$, $\theta = \{P(z_k|d_i)\}$, $\varphi = \{P(z_k|u_l)\}$ and $\psi = \{P(z_k|v_m)\}$. When a probabilistic model involves unobserved latent variables, the standard way for the maximum likelihood estimation of the model is the Expectation Maximization (EM) algorithm [9], which alternates two steps, E-step and M-step.

Let us first consider the special case that $\lambda = 0$. In such a case, the objective function boils down to the log-likelihood function of PLSA with no regularization term. Formally, we have the **E-step** to compute the posterior probabilities $P(z_k|d_i, w_j)$:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k'=1}^{K} P(w_j|z_{k'})P(z_{k'}|d_i)}. \quad (11)$$

In the **M-step**, we maximize the expected complete data log-likelihood for PLSA, which can be derived as:

$$Q_D = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} P(z_k|d_i, w_j) \log(P(w_j|z_k)P(z_k|d_i)).$$

There is a closed-form solution to maximize Q_D:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{j'=1}^{M} \sum_{i=1}^{N} n(d_i, w_{j'}) P(z_k|d_i, w_{j'})}, \quad (12)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{j'=1}^{M} n(d_i, w_{j'})}. \quad (13)$$

However, there is no closed-form solution in the M-step for the general case $\lambda > 0$ with the complete likelihood function:

$$\mathcal{Q}(\phi, \theta, \varphi, \psi) = Q_D - \lambda R(G).$$

Fortunately, we could use the generalized EM algorithm [19, 17] to maximize the regularized log-likelihood of the model. The major difference between generalized EM and traditional EM is that in the M-step generalized EM finds parameters that only "improve" the expected value of the log-likelihood function rather than maximizing it. It is easy to see that our biased regularization model shares the same hidden variables with PLSA, and has exactly the same E-step as that of PLSA. Since the regularization $R(G)$ does not involve the parameters $P(w_j|z_k)$, we can get the same M-step estimation equation for $P(w_j|z_k)$ as in Eq. (12).

In the following, we discuss how to estimate the parameter values $\theta = \{P(z_k|d_i)\}$ as well as $\varphi = \{P(z_k|u_l)\}$ and $\psi = \{P(z_k|v_m)\}$. Let us first find $\theta_{t+1}^{(1)}$ using Eq. (13) which maximizes Q_D instead of $\mathcal{Q}(\phi, \theta, \varphi, \psi)$. We then try to start from $\theta_{t+1}^{(1)}$ and decrease $R(G)$, which can be done through Newton-Raphson method [20, 6]. Given a function $f(x)$ and the initial value x_t, the Newton-Raphson updating formula to decrease $f(x)$ is defined as $x_{t+1} = x_t - \gamma \frac{f'(x)}{f''(x)}$, where $0 \leq \gamma \leq 1$ is the step parameter. With $\theta_{t+1}^{(1)}$ and put $R(G)$ into the Newton-Raphson formula, we can decrease $R(G)$ by updating $P(z_k|d_i)$ in each step:

$$\begin{aligned} P(z_k|d_i)_{t+1}^{(n+1)} &= (1-\gamma) P(z_k|d_i)_{t+1}^{(n)} \\ &+ \frac{\gamma}{2} \left(\sum_{u_j \in \mathcal{U}_{d_i}} \frac{P(z_k|u_j)}{|\mathcal{U}_{d_i}|} + \sum_{v_j \in \mathcal{V}_{d_i}} \frac{P(z_k|v_j)}{|\mathcal{V}_{d_i}|} \right). \end{aligned} \quad (14)$$

In the meantime, $P(z_k|u_l)$ and $P(z_k|v_m)$ are updated as in Eq. (7) and Eq. (9), respectively, for each step. The step parameter γ in Eq. (14) can be interpreted as a controlling factor of smoothing the topic distribution among the neighboring objects. We repeatedly update $\theta_{t+1}^{(n)}$ using Eq. (14) until $\mathcal{Q}(\theta_{t+1}^{(n+1)}) \leq \mathcal{Q}(\theta_{t+1}^{(n)})$. Then we test whether $\mathcal{Q}(\theta_{t+1}^{(n)}) \geq \mathcal{Q}(\theta_t)$. If it is true, re-estimation for θ is done by setting $\theta_{t+1} \leftarrow \theta_{t+1}^{(n)}$. Otherwise, we keep current θ, φ and ψ without updating in the M-step and continue to the next E-step. We summarize the model fitting approach by using generalized EM algorithm in Algorithm 1.

Algorithm 1 Model Fitting for Biased Regularization

Input: A text-rich information network $G = (D \cup U \cup V, E)$ with word occurrences $n(d_i, w_j)$. The number of topics K, Newton step size γ, regularization parameter λ
Output: $\phi = \{P(w_j|z_k)\}$, $\theta = \{P(z_k|d_i)\}$, $\varphi = \{P(z_k|u_l)\}$ and $\psi = \{P(z_k|v_m)\}$.

```
Random initialize the probability distribution φ_0 and θ_0, compute ϕ_0 = {P(z_k|u)_0} and ψ_0 = {P(z_k|v)_0} as in Eq. (7) and Eq. (9), respectively;
t ← 0;
while t < MaxIteration do
    E-step: Compute P(z_k|d_i, w_j) as in Eq. (11);
    M-step:
    Re-estimate P(w_j|z_k)_{t+1} as in Eq. (12);
    Re-estimate P(z_k|d_i)_{t+1} as in Eq. (13);
    P(z_k|d_i)^{(1)}_{t+1} ← P(z_k|d_i)_{t+1};
    Compute P(z_k|d_i)^{(2)}_{t+1} (i.e., θ^{(2)}_{t+1}) as in Eq. (14), and update P(z_k|u)_{t+1} and P(z_k|v)_{t+1} as in Eq. (7) and Eq. (9), respectively;
    while Q(θ^{(2)}_{t+1}) ≥ Q(θ^{(1)}_{t+1}) do
        P(z_k|d_i)^{(1)}_{t+1} ← P(z_k|d_i)^{(2)}_{t+1};
        Compute θ^{(2)}_{t+1}, update P(z_k|u)_{t+1} and P(z_k|v)_{t+1};
    end while
    if Q(θ^{(1)}_{t+1}) ≥ Q(θ_t) then
        P(z_k|d_i)_{t+1} ← P(z_k|d_i)^{(1)}_{t+1};
        Update P(z_k|u)_{t+1} and P(z_k|v)_{t+1};
    else
        Keep current θ, φ and ψ.
    end if
    t ← t + 1
end while
```

5. EXPERIMENTS

In this section, we evaluate the effectiveness of our TMBP algorithm, and compare it with the state-of-the-art methods on two data sets through extensive experiments.

5.1 Data Collection

The Digital Bibliography and Library Project (DBLP)[1] is a collection of bibliographic information on major computer science journals and proceedings, which can be used to build a heterogeneous information network with multi-typed objects along with rich text data as Figure 1(a). Each paper is represented by a bag of words that appeared in the abstract and title of the paper. Besides the rich-text documents, we also obtain two other types of objects: author and venue (i.e., conference). In this experiment, we use a subset of the DBLP records[2] that belongs to four areas: *database, data mining, information retrieval* and *artificial intelligence*, and contains 28,569 documents, 28,702 authors and 20 conferences. The abstract is collected for representing each document, and this data collection has 11,771 unique terms. Within the heterogeneous information network, we observe two explicit types of relationships: paper-author and paper-venue, which consist of a total number of 103,201 links. Moreover, we use a labeled data set [24] with 4,057 authors, 100 papers and all 20 conferences for quantitative accuracy evaluation.

The NSF Research Awards Abstracts (NSF-Awards)[3] consists of 129,000 abstracts describing NSF awards for basic research from 1990 to 2003, which are grouped into more than 640 research programs. For each NSF award, we obtain the abstract represented by a bag of words, and the affiliated investigator(s), forming a heterogeneous information network. In our test, we extract a subset of documents

[1] http://www.informatik.uni-trier.de/~ley/db/

[2] http://www.cs.uiuc.edu/~hbdeng/data/kdd2011.htm

[3] http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html

Table 1: Statistics of the DBLP and NSF datasets.

	DBLP	NSF-Awards
# of docs (D)	28,569	16,405
# of authors/PIs (U)	28,702	9,989
# of venues (V)	20	-
# of links (D-U)	74,632	20,717
# of links (D-V)	28,569	-
# of terms	11,771	18,674
# of clusters (K)	4	10

that belong to the largest 10 research programs, such as *applied mathematics, economics* and *geophysics*, thus leaving us with 16,405 documents and 9,989 associated investigators. Within the heterogeneous information network, there are a total of 20,717 links between documents and investigators. Moreover, this data collection has 18,674 unique terms which appear in all the abstracts. Table 1 provides the statistics of these two datasets. Note that we set the number of topics (K) to be 4 and 10 for DBLP and NSF-Awards, respectively.

5.2 Experimental Setup and Metrics

The proposed TMBP algorithm can be applied to different text mining tasks, such as topic modeling and object clustering. We evaluate the performance of our models in two frameworks: the biased random walk framework (TMBP-RW) and biased regularization framework (TMBP-Regu). For further performance comparison, we implemented other state-of-the-art methods as follows:

- Nonnegative Matrix Factorization (NMF) [16]
- Probabilistic Latent Semantic Analysis (PLSA) [13]
- Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [6]
- Latent Dirichlet allocation (LDA) [3]
- Author-Topic Model (ATM) [22]
- Ranking-based Clustering (NetClus) [24].

Since NMF, PLSA and LDA cannot be directly applied to heterogeneous information networks, only documents are utilized for these models. For LapPLSI, we constructed a homogeneous nearest-neighbor graph, and empirically set the number of nearest neighbors to 10, and the step parameter γ to 0.1. Moreover, the regularization parameter was tuned to produce the best performance among 10, 100 and 1000. For NetClus, we implemented a topic-based NetClus algorithm which utilizes the topic distribution (obtained using PLSA) instead of the word distribution for each document. All the other parameter settings were set to be identical to TMBP.

To quantitatively compare TMBP with these methods, we adopt two popular metrics, accuracy (AC) and normalized mutual information (NMI) [27], to measure the clustering performance. The AC is definded as $AC = \frac{\sum_{i=1}^{n} \delta(a_i, map(l_i))}{n}$, where n denotes the total number of objects, $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $map(l_i)$ is the mapping function [6] that maps each cluster label l_i to the equivalent label from the data corpus. On the other hand, given the two sets of document clusters C and C', their mutual information metric $MI(C, C')$ is defined as: $MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}$, where $p(c_i)$ and $p(c'_j)$ denote the probabilities that a document arbitrarily selected from the corpus belongs to the clusters c_i and c'_j, respectively, and $p(c_i, c'_j)$ denotes the joint probability that arbitrarily selected document belongs to the clusters c_i as well as c'_j at the same time. Suppose $H(C)$ and $H(C')$ are the entropies of C and C', respectively, it reaches the maximum, $\max(H(C), H(C'))$, when the two sets of clusters are identical, whereas it becomes zero when the two sets are independent. In our experiments, we use the normalized mutual information NMI as the $MI(C, C')$ normalized by $\max(H(C), H(C'))$ which ranges from 0 to 1.

5.3 Experimental Results

We consider the question whether our proposed method can boost the performance of topic modeling and object clustering using the biased propagation algorithm. First, the experiments are performed to compare the task of object clustering with quantitative analysis. Then we investigate the parameter setting of our TMBP model. Finally, we analyze the topic modeling with some case studies.

5.3.1 Quantitative Analysis

For quantitative evaluation, we apply our models on the task of object clustering using both DBLP and NSF-Awards datasets. The hidden topics extracted by the topic modeling approaches can be regarded as clusters. The estimated conditional probability (e.g., $P(z_k|d_i)$ and $P(z_k|u_l)$) is used to infer the cluster label for each object. The clustering result is evaluated by comparing the cluster label of each object with its label provided by the data corpus.

Table 2 shows the clustering performance of different methods. For each method, 20 test runs were conducted, and the final performance scores were obtained by averaging the scores from the 20 tests. To make the comparison fair, we used the same random starting points for NMF, PLSA, LapPLSI, NetClus and TMBP. The italic results of PLSA in Table 2(a) is obtained by the special case of TMBP-RW with $\xi = 1$, which is set as our **baseline** model.

From Table 2, we observe that PLSA outperforms NMF on paper without using any network information. As expected, LapPLSI outperforms PLSA slightly by incorporating a homogeneous nearest-neighbor graph, and ATM outperforms LDA by considering the paper-author graph. However, both PLSA and LapPLSI fail to outperform TMBP (both TMBP-RW and TMBP-Regu) as well as ATM and NetClus. One reason is that ATM, NetClus and TMBP take into account the heterogeneous information network directly.

For DBLP, our TMBP approach simultaneously clusters all types of objects in different groups by considering both the text information and the heterogeneous information network. As we can see, both TMBP-RW and TMBP-Regu get significantly better performance than **baseline** PLSA, especially on the types of paper and author. Moreover, they can even achieve better results than the state-of-the-art ATM and NetClus algorithms. This shows that by considering the biased propagation on the heterogeneous information network and integrating with topic modeling, TMBP can have better topic modeling power for clustering objects.

For NSF-Awards, we only show the results for documents as there is no available label information for investigators. In general, we can observe similar results as DBLP. The results of TMBP-Regu and ATM are comparable, which outperform all the other methods. Additionally, the improvement of TMBP-Regu over other methods is more significant on the DBLP corpus than the NSF-Awards corpus. One possi-

Table 2: Object clustering performance of different methods on (a) DBLP and (b) NSF-Awards datasets.

(a) DBLP

Object	Paper (%)		Author (%)		Venue (%)		Average (%)	
Metric	AC	NMI	AC	NMI	AC	NMI	AC	NMI
NMF	44.55	22.92	-	-	-	-	44.55	22.92
PLSA	59.45	32.75	*65.0*	*37.97*	*80.0*	*74.74*	68.15	48.49
LapPLSI	61.35	33.93	-	-	-	-	60.70	33.37
LDA	47.00	20.48	-	-	-	-	47.00	20.48
ATM	77.00	52.21	74.13	40.67	-	-	75.57	46.44
NetClus	65.00	40.96	70.82	47.43	79.75	76.69	71.86	55.03
TMBP-RW	73.10	53.13	82.59	67.76	81.75	**77.53**	79.15	66.14
TMBP-Regu	**79.15**	**59.16**	**89.81**	**74.25**	**82.75**	76.56	**83.90**	**69.99**

(b) NSF-Awards

Object	Doc (%)	
Metric	AC	NMI
NMF	45.97	40.92
PLSA	63.00	64.48
LapPLSI	63.65	64.58
LDA	65.06	63.36
ATM	65.69	69.58
NetClus	63.51	66.11
TMBP-RW	64.84	68.74
TMBP-Regu	65.15	69.83

ble reason is that the heterogeneous information network of NSF-Awards is much sparser than that of DBLP, in which there are only 1.26 links per document for NSF-Awards, and 3.61 links per document for DBLP. Although the link information is very limited in NSF-Awards, our approach can still improve the performance over baseline methods which confirms its effectiveness.

By comparing the results of TMBP-Regu with TMBP-RW, it is obvious that TMBP-Regu performs better than TMBP-RW. The improvement over the biased random walk method owes to the direct optimization of the heterogeneous information analysis and topic modeling in a unified regularization framework. This observation supports the theoretical analysis of our biased regularization framework that can provide valuable and reinforced information as well as the constraints of overfitting for topic modeling.

5.3.2 Parameter Analysis

In our method, there are two essential parameters, the biased parameter ξ for TMBP-RW and the regularization parameter λ for TMBR-Regu. In this subsection, the effect of parameters ξ and λ is studied and evaluated.

Figure 3 shows how the performance of TMBP-RW varies with the biased parameter ξ. As mentioned before, the biased parameter is used to control the balance between inherent topic distribution and the propagated topic distribution. When $\xi = 1$, it is the **baseline** PLSA model. We can see that the performance is improved over the baseline when incorporating the random walk on the heterogeneous network with $\xi < 1$. The changes are relatively small in Figure 3 (c) and (d) since the topic propagation is constrained by the limited links. With the decrease of ξ, the performance becomes worse, and even worse than the baseline, as the model relies more on the topic consistency while ignores the intrinsic topic of the documents. We empirically set the biased parameter $\xi = 0.9$ in other experiments.

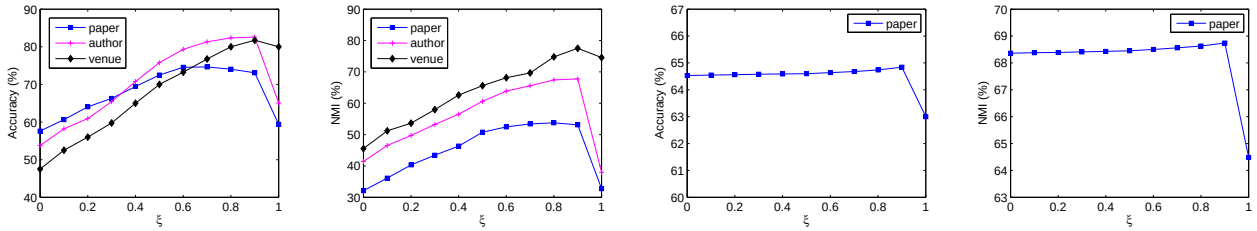
Figure 4 shows how the performance of TMBR-Regu varies with the regularization parameter λ. As mentioned in Section 4.2, the parameter λ is used to control the trade-off between the data likelihood of the topic modeling and the smoothness of topic distributions over the heterogeneous network. When $\lambda = 0$, the regularization framework boils down to be the baseline PLSA model. When $\lambda > 0$, the regularization framework takes into account the topic consistency between documents and their associated objects. As we can see, the TMBP-Regu is relatively stable with respect to the parameter λ, and achieves consistent good performance varying from 400 to 4000. We empirically set the biased parameter $\lambda = 1000$ in other experiments.

Table 3: The representative terms generated by PLSA, ATM and TMBP-Regu models. The terms are selected according to the probability $P(w|z)$.

Topic 1 (DB)	Topic 2 (DM)	Topic 3 (IR)	Topic 4 (AI)
PLSA			
data	data	information	problem
database	mining	retrieval	algorithm
systems	learning	web	paper
query	based	based	reasoning
system	clustering	*learning*	logic
databases	classification	knowledge	based
management	algorithm	text	time
distributed	*image*	search	algorithms
queries	analysis	system	*search*
relational	detection	language	show
ATM			
data	*learning*	information	knowledge
database	data	web	based
query	mining	retrieval	model
systems	algorithm	search	problem
databases	clustering	based	reasoning
queries	based	text	logic
system	classification	language	image
processing	algorithms	user	system
distributed	time	semantic	recognition
management	analysis	document	representation
TMBP-Regu			
data	data	information	learning
database	mining	web	based
query	algorithm	retrieval	knowledge
databases	clustering	search	model
systems	classification	based	problem
queries	based	text	reasoning
system	algorithms	language	system
processing	rules	user	logic
management	analysis	semantic	image
distributed	discovery	document	models

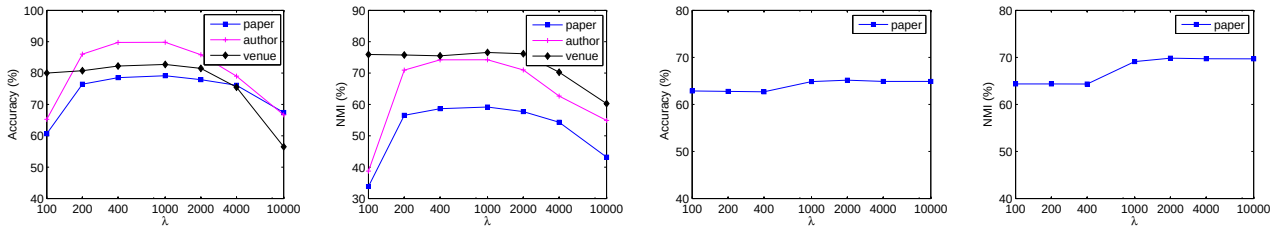
5.3.3 Topic Modeling Analysis and Case Study

In order to visualize the hidden topics and compare different approaches, we extract topics from the data using PLSA, ATM and TMBP-Regu on DBLP dataset. Since the DBLP subset is a mixture of four areas, it is interesting to see whether the extracted topics could automatically reveal this mixture. The most representative terms generated by PLSA, ATM and TMBP-Regu are shown in Table 3. For the first three topics, although different algorithms select slightly different terms, all these terms can describe the corresponding topic to some extent. For Topic 4 (AI), the top keywords like "learning, based, knowledge" derived from TMBP-Regu is obviously more telling than "knowledge, based, model" derived by ATM and "problem, algorithm, paper" derived by PLSA. Similar subtle differ-



(a) DBLP (b) DBLP (c) NSF-Awards (d) NSF-Awards

Figure 3: The effect of varying parameters ξ in the biased random walk framework (TMBP-RW).



(a) DBLP (b) DBLP (c) NSF-Awards (d) NSF-Awards

Figure 4: The effect of varying parameters λ in the biased regularization framework (TMBP-Regu).

ences can be observed for Topic 2 (DM) as well. Intuitively, TMBP-Regu and ATM select more related terms for each topic than PLSA, which shows the better performance of TMBP-Regu and ATM by considering the heterogeneous information network.

5.3.4 *Summary*

These experimental results demonstrate the effectiveness of our proposed algorithm, which successfully incorporates the heterogeneous information network into topic modeling. Compared with PLSA, TMBP-RW keeps learned topics the same as PLSA, and propagates topic probabilities biasedly between documents and other objects. In contrast, TMBP-Regu results in refined topic-word distribution, and derives the topic probabilities for all types of objects collaboratively in a unified way, leading to better performance as expected. In comparison with ATM, TMBP-Regu is more flexible to deal with all types of objects including authors, and more straightforward to incorporate additional graph information.

6. RELATED WORK

Many topic models have been proposed and shown to be useful for data analysis. There are two principal approaches, PLSA [13] and LDA [3], which have been successfully applied or extended to many problems, including document clustering and classification [6, 7, 14], information retrieval [26, 30], correlated and dynamic topic models [1, 2], geographical topic discovery [28], author-topic modeling [22, 25], and citation and social network analysis [17, 6, 8, 18]. However, most of these models only consider the textual information while ignore network structures. Recently, several studies, including NetPLSA [17], LapPLSI [6], LTM [7] and iTopicModel [23], have been proposed for combining topic modeling with homogeneous networks, such as citation graph and co-authorship graph, but they cannot deal with heterogeneous information network directly. Although there was some research done to model the relationships between different objects, such as Author-Topic Model [22, 25] and collective topic model [12], these models are designed specifically for academic networks or only consider the relationships indirectly through the content information. Our proposed models differs from them as we directly take into account the general heterogeneous information networks with topic propagation techniques.

Link analysis has been a hot topic for a few years since the advent of two distinct methods, HITS [15] and PageRank [5]. Many techniques have been proposed to improve search results [4, 29] and integrate heterogeneous networks [11, 24]. For example, Sun et al. [24] proposed a ranking-based clustering for heterogeneous information networks. Deng et al. [11] developed a generalized Co-HITS algorithm for bipartite graph analysis. However, their algorithms are based on a simple and unbiased propagation method, and ignore the underlying latent topics associated with the network in most cases.

This work is also related to graph-based semi-supervised learning [33, 31, 21, 32], which usually assumes label smoothness over the graph. These types of graph regularization methods have been successfully applied in many data mining tasks [17, 6, 10]. In [10], the authors developed a graph-based re-ranking model by regularizing the smoothness of relevance scores over the latent graph. NetPLSA [17] and LapPLSI [6] explored graph-based regularizers with topic modeling by considering homogeneous networks. Our work is different from theirs, as we focus on heterogeneous information networks and introduce a new biased regularization term, which distinguishes documents with rich text and other objects without explicit text, so as to treat them in a different way.

7. CONCLUSIONS AND FUTURE WORK

We have presented a novel algorithm for topic modeling on text-rich heterogeneous information networks, called Topic Model with Biased Propagation (TMBP). Consequently, we have investigated the biased random walk (TMBP-RW) and biased propagation (TMBP-Regu) frameworks, for incorpo-

rating heterogeneous information network into topic modeling directly. As a result, TMBP can make full use of both rich semantics embedded in heterogeneous networks and rich text of documents, which leads to a significant improvement over the baseline topic models. Moreover, TMBP-Regu performs better than TMBP-RW since topic modeling and heterogeneous network analysis can mutually enhance each other in the biased regularization framework. Experimental results on object clustering and topic modeling show that TMBP-Regu achieves the best performance. In future work, the idea of the biased propagation framework can also be naturally incorporated with other topic modeling algorithms, e.g., Latent Dirichlet Allocation. It would be interesting to investigate the performance of our algorithm by varying the weights of different objects besides documents.

8. ACKNOWLEDGMENTS

The work was supported in part by the U.S. National Science Foundation grants IIS-0905215, CNS-0931975, by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), and by the U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

9. REFERENCES

[1] D. M. Blei and J. D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *TOIT*, 5(1):231–297, 2005.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[6] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM*, pages 911–920, 2008.

[7] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, page 14, 2009.

[8] D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, pages 430–436, 2000.

[9] A. Dempster, N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[10] H. Deng, M. R. Lyu, and I. King. Effective latent space graph-based re-ranking model with global consistency. In *WSDM*, pages 212–221, 2009.

[11] H. Deng, M. R. Lyu, and I. King. A generalized Co-HITS algorithm and its application to bipartite graphs. In *KDD*, pages 239–248, 2009.

[12] H. Deng, B. Zhao, and J. han. Collective topic modeling for heterogeneous networks. In *SIGIR*, 2011.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[14] S. Huh and S. E. Fienberg. Discriminative topic modeling based on manifold learning. In *KDD*, pages 653–662, 2010.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[17] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, pages 101–110, 2008.

[18] R. Nallapati, A. Ahmed, E. P. Xing, and W. Cohen. Joint latent topic models for text and citations. In *KDD*, pages 542–550, 2008.

[19] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, 89:355–368, 1998.

[20] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. Numerical Recipes in C: The Art of Scientific Computing, Cambridge, 1992.

[21] A. Smola and R. Kondor. Kernels and regularization on graphs. *COLT*, 2003.

[22] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. Probabilistic author-topic models for information discovery. In *KDD*, pages 306–315, 2004.

[23] Y. Sun, J. Han, J. Gao, and Y. Yu. itopicmodel: Information network-integrated topic modeling. In *ICDM*, pages 493–502, 2009.

[24] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*, pages 797–806, 2009.

[25] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998, 2008.

[26] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[27] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.

[28] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.

[29] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR*, pages 504–511, 2005.

[30] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *WWW*, pages 715–724, 2008.

[31] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.

[32] D. Zhou, B. Schölkopf, and T. Hofmann. Semi-supervised learning on directed graphs. In *NIPS*, 2004.

[33] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.