iTopicModel: Information Network-Integrated Topic Modeling

Yizhou Sun, Jiawei Han, Jing Gao and Yintao Yu
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
{*sun22,hanj,jinggao3,yintao*}*@illinois.edu*

***Abstract*—*Document networks*, *i.e.*, networks associated with text information, are becoming increasingly popular due to the ubiquity of Web documents, blogs, and various kinds of online data. In this paper, we propose a novel topic modeling framework for document networks, which builds a unified generative topic model that is able to consider both text and structure information for documents. A graphical model is proposed to describe the generative model. On the top layer of this graphical model, we define a novel multivariate Markov Random Field for topic distribution random variables for each document, to model the dependency relationships among documents over the network structure. On the bottom layer, we follow the traditional topic model to model the generation of text for each document. A joint distribution function for both the text and structure of the documents is thus provided. A solution to estimate this topic model is given, by maximizing the log-likelihood of the joint probability. Some important practical issues in real applications are also discussed, including how to decide the topic number and how to choose a good network structure. We apply the model on two real datasets, DBLP and Cora, and the experiments show that this model is more effective in comparison with the state-of-the-art topic modeling algorithms.**

***Keywords*-document networks; topic model; Markov Random Field.**

I. Introduction

Document networks, *i.e.*, information networks associated with text information, are ubiquitous and indispensable nowadays due to the popular use of web, blogs, and various kinds of online databases. Examples of document networks are: co-author networks and citation networks with text extracted from publications for each author/paper in bibliographic databases like DBLP[1]; social network with text extracted from blogs and posts for each user in social network sites like Facebook[2]; actor cooperation network with text as movie plots they have stared in movie databases like IMDB[3], and so on. In this paper, we study the problem of building topic models on arbitrary document networks, either weighted or unweighted, directed or undirected.

Similar to traditional topic modeling methods, such as PLSA [1] and LDA [2], topic modeling on document networks tries to soft clustering the documents into different clusters with the meaning of topics, and each topic is described using a multinomial distribution over words. Thus, each topic can be easily understood by browsing only several top probability words in the distribution. Moreover, by presenting topic membership probabilities (referred as *topic distribution* thereafter) for each document, people may understand the general content of that document. In traditional topic modeling methods, documents are assumed independent with each other, and no links among them will be considered in the modeling process. However, in real life, two documents can be linked together through all sorts of semantics. For example, two papers can be linked together via citations, two webpages can be linked together by their hyper links, and two authors can be linked together according to the co-author relationship. More importantly, intuitively, two closely related documents should have similar text information, which can be utilized to improve the topic modeling. For example, if two researchers co-author a lot, we can infer that they share similar topics. A set of ideal independent documents in traditional topic model, and a set of mutually dependent (or connected) documents in real case are illustrated in Figure 1 (a) and (b), respectively.

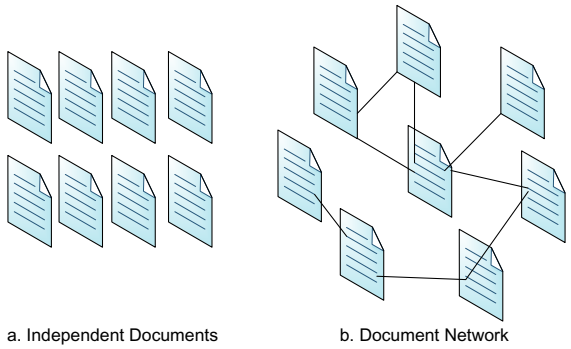

Figure 1. Two views on documents sets: Traditional topic model vs. iTopicModel.

The work was supported in part by MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC, the U.S. National Science Foundation grants IIS-08-42769 and IIS-0905215, and the Air Force Office of Scientific Research MURI award FA9550-08-1-0265. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

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

[2]http://www.facebook.com

[3]http://www.imdb.com

In the studies of information networks, clustering on networks [3] or graph partitioning [4], [5], has been a popular topic for over a decade. Similar to our problem setting, similar objects are grouped into clusters according to links among them. Unfortunately, most of these studies are unable to utilize the text information that each node may contain. Thus, clusters so detected may not be quite accurate, especially when the network structure is sparse, which is often the case in real life. Moreover, since the clusters can only be described using objects rather than text, it is not easy for users to understand the semantic meaning of clusters so obtained, especially when the network is large. Another weak point for merely using link information of a network in clustering is that, connectivity is often required for the network, otherwise an outlier node or an outlier node group is either partitioned into a separate cluster, or randomly assigned to one cluster. However, for example, in a co-author network, a small group of authors may never co-author with other authors, but we would still be interested in their research areas.

Obviously, if one can utilize both the link information among documents and text information for each document, the topic modeling will be significantly improved. Currently, there are some approaches proposed to improving traditional topic models by integrating the complex structural information with text. One is to design complex generative model for text considering the semantic meaning associated with links [6], [7]. Another is to add a graph regularization constraint to the original log-likelihood objective function [8]. For the first approach, the complex generative model requires lots of expert knowledge and may not be easy to migrate to other datasets; whereas for the second, the combined objective function requires a parameter to adjust the weight of two original objective functions, and it lacks a unified generative model explanation.

In this paper, we propose a unified generative model considering both text and structure information in a document network, *i.e.*, an information network with each node containing text information. According to this model, each topic is modeled with a multinomial distribution over words, and each document in the network is associated with a T dimensional random variable $\mathbf{\Theta_i}$, representing a topic distribution vector, and the dependency relationships among the variables are modeled using a multivariate Markov Random Field, given the network structure. Based on the generative model for the text and the structure, we then give the parameter estimations by maximizing the log-likelihood of the joint distribution of the generative model. Moreover, we propose a Q-function-based topic number selection method, which can help users decide the best topic number T, given the current network with text. Thus, the contributions of this study can be summarized as follows.

- A unified generative topic model is proposed for document networks which considers both text and structure information, and an efficient solution is provided to estimate its parameters;
- a method is proposed to determine the topic number by utilizing the link information in the network; and
- experiments are conducted on two real datasets, and the results show the effectiveness of our model in comparison with the existing state-of-the-art models.

The remaining of the paper is organized as follows. In Section 2, we give a brief introduction to the related work. In Section 3, we introduce our unified generative model and give the estimation formulas to its parameters. In Section 4, we discuss several related issues in building generative model for different real datasets for practical applications. Section 5 is the experiment study, which compares our model with several state-of-the-art models in two real datasets. Section 6 concludes this study.

II. Related Work

Topic modeling over documents is to find T topics that best describe the given corpus by assuming each document is a mixture model of these topics, where T is given and each topic is described using a multinomial distribution over words. PLSA (Probabilistic Latent Semantic Analysis) [1] and LDA (Latent Dirichlet Allocation) [2] are two most well known topic modeling methods. However, both PLSA and LDA treat documents in a given corpus as independent to each other. For example, in LDA, the random variable of topic distribution parameter for each document $\mathbf{\Theta}_i$ is assumed i.i.d from a Dirichlet distribution. However, the independence assumption may not hold in real cases, especially when documents are linked to each other via links in the networks.

In order to integrate the structure information of documents into topic modeling, there are three lines of studies. The first is to build complex generative models for documents, given the additional structural information. For example, in author-topic model [6], [7], a document is modeled as first choosing an author, then selecting a topic according to the specific author's topic distribution, and then selecting a word from the corresponding topic. Such methods require expert knowledge on the semantic meanings of the links and is difficult to migrate to other datasets. The second line of study is to add a regularization constraint on networks to the traditional topic models, such as in the recently proposed NetPLSA [8]. This type of methods combines two objective functions into a new objective function, but lacks a generative explanation to such combinations. Also, NetPLSA can only deal with undirected networks. A newly proposed method called Relational Topic Model (RTM) [9] proposes another view and tries to model nodes and links separately. However, RTM can only model unweighted networks, where a link is either observed or unobservable. Different from the three methods, in this paper, we propose a *unified* generative model integrating both text information

and structural information, which is able to be applied to *any* document networks, either weighted or unweighted, directed or undirected. Other related studies include document classification and clustering that integrate both text and structural information, such as in [10], [11], however, they have not addressed the topic modeling problem, but aim at extracting good features for better clustering or classification.

Clustering on networks, which aims at clustering nodes of the network into different groups, has been studied extensively. The most well known family of such methods could be spectral clustering methods [5], and NCut [4] is one of the most popular criteria in such algorithms. However, such algorithms do not consider text information associated with each node to help clustering.

Markov Random Field (MRF) [12], [13], [14] provides a way to model the dependency among random variables, according to their structural information described in a graph. MRF has many applications in image processing, spatial data analysis, and so on. In this paper, we will define a novel multivariate MRF over the random variables of topic distribution for each document, and model their dependency using the links in the document network.

III. Modeling for iTopicModel

In this section, we build a unified generative model by integrating both structural and text information in a document network.

A. Preliminaries

We first define some terms and notations that will be used in the following context.

Definition 1. ***Document****. A document x_i in a document collection $X = \{x_1, x_2, \ldots, x_N\}$ is comprised of a bag of words from a vocabulary $Y = \{y_1, y_2, \ldots, y_M\}$, and is represented with vector $\mathbf{x}_i = (c_{i1}, c_{i2}, \ldots, c_{iM})$, where c_{il} denotes the occurrence number of word y_l in document x_i.*

Definition 2. ***Document Network****. A document network $G = \langle X, E, W\rangle$ is a graph defined on a document set X. E is the link set, and $e = \langle x_i, x_j\rangle \in E$ if there is a link from document x_i to x_j. W is the adjacency matrix denoting the weights of the links, $w_{ij} > 0$ if there is a link from node x_i to x_j, and the value of w_{ij} is the strength of the link $e = \langle x_i, x_j\rangle$; $w_{ij} = 0$, otherwise.*

Definition 3. ***Neighborhood****. The neighborhood of a given document x_i in the document network G, denoted as $N(i)$, is defined as $N(i) = N_{out}(i) \cup N_{in}(i)$, where $N_{out}(i) = \{x_j | \langle x_i, x_j\rangle \in E\}$ and $N_{in}(i) = \{x_j | \langle x_j, x_i\rangle \in E\}$, representing the out-neighborhood and in-neighborhood respectively.*

In this paper, we confine our study on document networks with nonnegative weights on links. For undirected networks, they will be transformed to directed networks by converting each undirected link into two directed links. For unweighted networks, the weights of links are defined either 1 or 0, representing the status of observed and unobserved respectively. An example of directed document network is given in Example 3.1, which is a paper citation network.

Example 3.1 (Paper Citation Network) Let $X = \{x_1, x_2, \ldots, x_N\}$ be the collection of all the papers in a bibliographic database, each paper x_i is a document, which is comprised of a bag of words from a vocabulary $Y = \{y_1, y_2, \ldots, y_M\}$. The text information for each paper can be from titles, abstracts, or even full text, according to the information availability of the database. For example, in DBLP, only titles are available for each paper, while for ACM Digital Library[4], abstracts can also be obtained. We build a network among these papers according to their citation relationship, *i.e.*, if x_i cites x_j, a link $e = \langle x_i, x_j\rangle$ with the weight $w_{ij} = 1$ is then added to E. ∎

For topic modeling, given the topic number T, each topic is modeled as a multinomial distribution over words, with the parameter $\boldsymbol{\beta}_{T\times M} = \{\beta_{kl}\}$ and $\sum_{l=1}^{M} \beta_{kl} = 1$, denoting the probability of word y_l in topic k. Each document then can be viewed as a mixture model over the T topics, and $\boldsymbol{\theta} = \{\theta_{ik}\}$ denotes the probability that x_i belongs to topic k, with the constraints that $\sum_{k=1}^{T} \theta_{ik} = 1$. Our topic modeling is to find the best $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ that maximizes the joint distribution of a document network given the current observation of text information and structural information.

For the self containment of this paper, we give a brief introduction to Markov Random Field, mainly following the work of [14].

Definition 4. ***Markov Random Field****. Given a graph $G = \langle V, E\rangle$, where $V = \{1, \ldots, n\}$, with each number as the label for each node. Let $F = \{F_i\}_{i=1}^{n}$ be a family of random variables defined on the node set V, i.e., each node i is associated with a random variable F_i. F is said to be a* ***Markov Random Field*** *on V with respect to graph G if and only if the following two conditions are satisfied:*

$$P(f) > 0, \ \forall f \in \mathbb{F} \tag{1}$$

$$P(f_i | f_{-i}) = P(f_i | f_{N(i)}) \tag{2}$$

where $f = \{f_1, \ldots, f_n\}$ is a configuration *of $F = \{F_i\}_{i=1}^{n}$, $P(f)$ is the abbreviation of $P(F = f)$, $P(f_i)$ is the abbreviation of $P(F_i = f_i)$, $\mathbb{F}$ is all the possible configuration set for F, $\{-i\}$ denotes the node set $V - \{i\}$, which is shortened as $-i$, $f_{-i} = \{f_j | j \neq i\}$ denotes the configuration of $F_{-i} = \{F_j | j \neq i\}$ defined on the nodes $V - \{i\}$, and $N(i)$ denotes the neighbors of the node i.*

When F_i is a multivariate variable, we call such MRF multivariate Markov Random Fields. Eq. (2) is called the

[4]http://portal.acm.org/dl.cfm

Notations	Meanings
Plain capital letter F	random variables
Plain small letter f	specific values for their random variables
Bold capital letters $\mathbf{\Theta}_i$	multivariate random variable associating with document x_i, indicating document topic distribution
Bold capital letters $\mathbf{\Theta} = \{\mathbf{\Theta}_i\}_{i=1}^N$	a random variable family, each of which is a multivariate random variable
Bold small letters $\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i$	a vector indicating topic distribution for document x_i, which is a specific value for random vector $\mathbf{\Theta}_i$
Bold small letters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}_{i=1}^N, \boldsymbol{\theta}' = \{\boldsymbol{\theta}'_i\}_{i=1}^N$	a configuration of $\mathbf{\Theta}$
Bold small letters $\boldsymbol{\beta}$	parameters indicating word distribution for topics
Plain capital Greek letter $\Psi = (\boldsymbol{\theta}, \boldsymbol{\beta})$	parameter for topic generative model
Bold small letters $\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^0$	Dirichlet distribution parameters and priors for $\mathbf{\Theta_i}$
Plain small letters i, j, k, l	indices: i, j for documents, k for topics, and l for words
Plain small letters with subscripts $\theta_{ik}, \beta_{kl}, \alpha_{ik}$	elements of matrix or vector

Table I
NOTATIONS

markovianity property of MRF, also called the *local property*. According to Hammersley-Clifford theorem, an MRF defined as above can be factorized into the form of $P(f) = \frac{1}{Z}\exp\{-U(f)\}$, where $Z = \sum_{f\in\mathbb{F}}\exp\{-U(f)\}$, is called the *partition function*, and $U(f) = \sum_{c\in\mathcal{C}} V_c(f)$, is called the *energy function*. $V_c(f)$ is defined over cliques c in the graph G, and is called *clique potentials*. Therefore, there are two equivalent methods to define an MRF: one is using the local property to specify conditional probabilities $P(f_i|f_{N(i)})$, and the other is to directly give the joint probability $P(f)$, according to its *global property*. Some major notations are summarized in Table I.

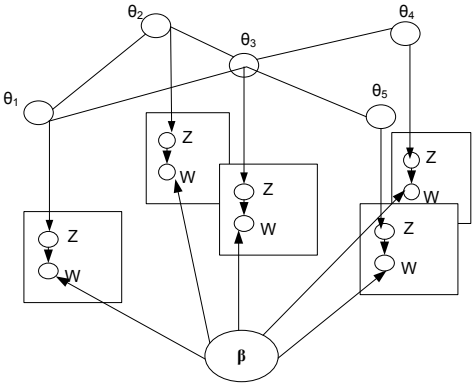

Figure 2. Graphical Model of iTopicModel

B. Model Set Up

To integrate structure information in the network into topic modeling, we now propose a novel unified generative model, iTopicModel, for generating documents, which has considered the dependency relationships among documents given by the document network. The graphical model of iTopicModel is in Figure 2, which can be viewed as two layers. The top layer has the same topology as the document network G. Let $\mathbf{\Theta}_i$ be the T dimensional multivariate random variable associated with document node x_i on G, and $\mathbf{\Theta}_i$ and $\mathbf{\Theta}_j$ be linked if and only if documents x_i and x_j are linked in network G. Let $\mathbf{\Theta}$ be $\{\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_N\}$, the family of multivariate random variables of $\mathbf{\Theta}_i$, and a multivariate Markov Random Field is defined on $\mathbf{\Theta}$ to model the dependency among documents, which will give different probabilities to different configurations of $\boldsymbol{\theta}$ for $\mathbf{\Theta}$. The bottom layer is composed of the traditional document generative models for each document, where each word is generated by first choosing a topic z with probability θ_{iz} according to $\boldsymbol{\theta}$, which is the current configuration of $\mathbf{\Theta}$, and then choosing a word y_l from the vocabulary following the distribution of topic z with β_{zl}. The joint probability for both text and structure of documents is then defined as:

$$p(X, \boldsymbol{\theta}|G, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|G)p(X|\boldsymbol{\theta}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|G)\prod_{i=1}^{N} p(x_i|\boldsymbol{\theta}_i, \boldsymbol{\beta}) \tag{3}$$

where documents are conditional independent with each other given current configuration of $\boldsymbol{\theta}$ for MRF $\mathbf{\Theta}$. We can see that the joint distribution for text and structure is decomposed into two parts, structure part denoted as $p(\boldsymbol{\theta}|G)$, and text generative part for each document x_i as $p(x_i|\boldsymbol{\theta}_i, \boldsymbol{\beta})$. We will give definitions for each part in the following.

1) Structure Modeling: Now we will define Markov Random Field $\mathbf{\Theta}$ on network G, and thus give the definition for the structure probability $p(\boldsymbol{\theta}|G)$. Intuitively, for each document, their topic distribution should be very similar to their neighboring documents. Therefore, we now try to model this similarity by specifying the probability of a configuration of a document using its neighbors' configurations. The higher probability the configuration is, the more similar the document and its neighbors are. For each variable $\mathbf{\Theta}_i$ associated with document x_i given its neighborhood, we

define it as a Dirichlet distribution, with the parameters derived from the out neighbor variables $\mathbf{\Theta}_{N_{out}(i)}$ and the weight of the links between them:

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{N_{out}(i)}) \sim Dirichlet(\boldsymbol{\alpha}_i) = \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{k=1}^{T} \theta_{ik}^{\alpha_{ik}-1} \quad (4)$$

where $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i|N_{out}(i)} = \boldsymbol{\alpha}_i^0 + \sum_{j \in N_{out}(i)} w_{ij}\boldsymbol{\theta}_j$, $\boldsymbol{\alpha}_i^0$ is a prior vector, and $B(\boldsymbol{\alpha}_i) = \frac{\prod_{k=1}^{T}\Gamma(\alpha_{ik})}{\Gamma\left(\sum_{k=1}^{T}\alpha_{ik}\right)}$ is a multinomial beta function. We give such definition with the following intuitions:

1) In this definition, we only consider the out-neighborhood of a document, for the reasons: (1) for undirected networks, out-neighborhood is the same as the whole neighborhood; and (2) for directed networks, out links are more trustable than in links. For example, in a citation network, a document x_i could be cited by many documents with all sorts of reasons, but most of its references should be similar to it. Also, if a network is transformed to a top-k network, where only the top-k highest weight links are preserved for each document, it is also more reliable to use the out-neighbors, *i.e.*, its top-k most linked documents.
2) A document's topic distribution configuration $\boldsymbol{\theta}_i$ should be very close to weighted mean of its neighbors' topic distribution configurations $\boldsymbol{\theta}_{N_{out}(i)}$. This is satisfied, since $E(\mathbf{\Theta}_i|\mathbf{\Theta}_{N_{out}(i)}) = \frac{\boldsymbol{\alpha}_i}{|\boldsymbol{\alpha}_i|}$, where $|.|$ is the 1-norm of vector, according to the property of Dirichlet distribution.
3) The precision ([15]) about how close a configuration $\boldsymbol{\theta}_i$ is around the mean, which is $|\boldsymbol{\alpha}_i| = \sum_{k=1}^{T}(\alpha_{i,k}^0 + \sum_{j \in N_{out}(i)} w_{ij}\theta_{jk}) = |\boldsymbol{\alpha}_i^0| + \sum_{j \in N_{out}(i)} w_{ij}$, is determined by the total out degree of this document ($\sum_{j \in N_{out}(i)} w_{ij}$). The more out degree of a document, the more information we know about this document, thus the more we can trust with its neighbors ($\mathbf{\Theta}_i$ has higher confidence around the mean of $\boldsymbol{\alpha}_i/|\boldsymbol{\alpha}_i|$).

The prior $\boldsymbol{\alpha}_i^0$ is the prior knowledge we know about document x_i. It can be viewed that, there are T additional documents in the network, with each document purely from one topic (say k_{th}), *i.e.*, the topic distribution having the value 1 at the k_{th} component and all zeros at the remaining components. Document x_i then has links to each of them, with the weight of α_{ik}^0. When setting all $\boldsymbol{\alpha}_i^0$ as $\vec{1}$, they can be viewed as a smoothing prior over the graph structure, *i.e.*, each document can be viewed as linking to the additional T documents with the weight of 1. This smoothing is especially useful if a document has no neighbors in the network, otherwise there will be computational problems. When setting all of $\boldsymbol{\alpha}_i^0$ as 0, it can be viewed as no prior on the network structure at all. Other settings can be viewed as a prior knowledge of topic distribution $\boldsymbol{\theta_i}$ on each document. In the following experiments, we set $\boldsymbol{\alpha}_i^0$ as $\vec{1}$.

Next, we will give an MRF definition over $\mathbf{\Theta}$, which is able to give the probability to a configuration over all the documents and reflect the intuitions stated above. According to the global property of MRF, the density function of $\mathbf{\Theta}$ can be factorized into the form:

$$p(\boldsymbol{\theta}) = \frac{1}{Z}\exp\{-\sum_{c \in \mathcal{C}} V_c(\boldsymbol{\theta})\} \quad (5)$$

where $Z = \sum_{\boldsymbol{\theta}} \exp\{-\sum_{c \in \mathcal{C}} V_c(\boldsymbol{\theta})\}$ is the partition function[5], and c stands for the clique in graph G. In the following, we only consider the potential functions defined on single nodes and single edges.

Theorem 1. *Given the potential functions defined on nodes and edges as:*

$$V_i(\boldsymbol{\theta}_i) = -(\boldsymbol{\alpha}_i^0 - \vec{1})^T \log(\boldsymbol{\theta}_i)$$
$$V_{i \to j}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \begin{cases} -(w_{ij}\boldsymbol{\theta}_j)^T \log(\boldsymbol{\theta}_i), & \textit{if } \langle x_i, x_j \rangle \in E; \\ 0, & \textit{otherwise.} \end{cases}$$

define joint distribution $p(\boldsymbol{\theta})$ over $\mathbf{\Theta}$ using the form of Eq. (5), then the joint distribution is

$$p(\boldsymbol{\theta}|G) = \frac{1}{Z}\exp\{\sum_i [(\boldsymbol{\alpha}_i^0 + \sum_{j \in N(i)} w_{ij}\boldsymbol{\theta}_j - \vec{1})^T \log(\boldsymbol{\theta}_i)]\} \quad (6)$$

and $\mathbf{\Theta}$ is an MRF.

Proof: Joint distribution of Eq. (6) can be directly obtained. To prove $\mathbf{\Theta}$ is an MRF, two conditions of MRF need to be validated. Eq. (1) is easy to check. Markovianity of MRF denoted by Eq. (2) is also easy to check: $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}) = p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{N(i)}) \propto \exp\{-V_i(\boldsymbol{\theta}_i) - \sum_{j \in N_{out}(i)} V_{i \to j}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - \sum_{j \in N_{in}(i)} V_{j \to i}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)\}$. ■

From Eq. (6), we can see that the joint distribution of $\mathbf{\Theta}$ can be viewed as $\frac{1}{Z}\prod_{i=1}^{N} B(\boldsymbol{\alpha}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{N_{out}(i)})$, where $\boldsymbol{\alpha}_i$ and $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{N_{out}(i)})$ follow the definition in Eq. (4). This means a high probability global configuration for all the documents is the configuration with good agreements in the local structures.

2) Text Modeling: Now, we will consider the text part $p(x_i|\boldsymbol{\theta}_i, \boldsymbol{\beta})$ of the joint distribution in Eq. (3). By assuming each document is a mixture model over T topics, *i.e.*,

$$p(y_l|x_i) = \sum_{k=1}^{T} p(z=k|x_i)p(y_l|z=k) = \sum_{k=1}^{T} \theta_{ik}\beta_{kl},$$

and each word is generated following multinomial distribution, the probability to generate document x_i given parameter $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is:

$$p(x_i|\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{l=1}^{M} p(y_l|x_i, \boldsymbol{\theta}, \boldsymbol{\beta})^{c_{il}} = \prod_{l=1}^{M} [\sum_{k=1}^{T} \theta_{ik}\beta_{kj}]^{c_{il}} \quad (7)$$

[5]We use sum ($\sum$) instead of integration ($\int$) over all the configuration of $\boldsymbol{\theta}$ for better understanding, though $\boldsymbol{\theta}$ is now in a continuous space.

Notice that $p(x_i|\boldsymbol{\theta}_i, \boldsymbol{\beta}) = p(x_i|\boldsymbol{\theta}, \boldsymbol{\beta})$, since x_i is only dependent on topic distribution of the same document $\boldsymbol{\theta}_i$. In traditional topic modeling methods, documents are then assumed to be independent to each other to derive the joint distribution. In PLSA [1], the joint probability is equivalent to the multiplication of probabilities of observing each document $\prod_{i=1}^{N} p(x_i|\boldsymbol{\theta}, \boldsymbol{\beta})$. In LDA [2], the distributions over topics for each document $\boldsymbol{\Theta}_i = (\Theta_{i1}, \ldots, \Theta_{iT})$ is assumed to follow the Dirichlet distribution with parameter $\boldsymbol{\alpha}$, which serves as a prior, independently, and the joint probability is the multiplication of the probabilities of generating each of the documents $\prod_{i=1}^{N} p(x_i|\boldsymbol{\alpha}, \boldsymbol{\beta})$.

C. Parameter Estimation

In Section III-B, a generative model for document network called iTopicModel has been proposed and a joint probability function of both text and structure is thus defined. In this section, we will give the parameter estimation method by maximizing the log-likelihood of the joint distribution.

$$\begin{aligned}\log L &= \log p(X, \boldsymbol{\theta}|G, \boldsymbol{\beta}) \text{ (replacing with Eqs. (6) and (7))} \\ &= \sum_{i=1}^{N}\sum_{k=1}^{T}\Big((\alpha_{i,k}^{0} - 1)\log\theta_{ik} + \sum_{j=1}^{N} w_{ij}\theta_{jk}\log\theta_{ik}\Big) \\ &+ \sum_{i=1}^{N}\sum_{l=1}^{M} c_{il}\log(\sum_{k=1}^{T}\beta_{kl}\theta_{ik}) - \log Z\end{aligned} \tag{8}$$

where Z is a constant and can be neglected for maximization.

The parameters we are going to estimate are $\Psi = (\boldsymbol{\theta}_{\{N\times T\}}, \boldsymbol{\beta}_{\{T\times M\}})$, which is the same as in PLSA and $NT + TM$ in total. We now use an approximate EM algorithm to get the best estimators by maximizing the log-likelihood in Eq. (8), with the constraints that $\sum_{k=1}^{T}\theta_{ik} = 1$ for all i and $\sum_{l=1}^{M}\beta_{kl} = 1$ for all k. The hidden variable in the likelihood function is the topic indicator variable z for each word in each document, and $p(z = k|x_i, y_l, \Psi) \propto p(z = k, y_l|x_i, \Psi) = p(z|x_i, \Psi)p(y_l|z, \Psi)$, thus:

$$p(z = k|x_i, y_l, \Psi) = \frac{\beta_{kl}\theta_{ik}}{\sum_{k'=1}^{T}\beta_{k'l}\theta_{ik'}} \tag{9}$$

In the E-step, conditional expectation of $\log L$ given current value of parameters and conditional distribution of z is:

$$\begin{aligned}Q(\Psi|\Psi^{(t)}) &= E_{z|X,\Psi^{(t)}}(\log L) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{T}\Big((\alpha_{i,k}^{0} - 1)\log\theta_{ik} + \sum_{j=1}^{N} w_{ij}\theta_{jk}^{(t)}\log\theta_{ik}\Big) \\ &+ \sum_{i=1}^{N}\sum_{l=1}^{M} c_{il}\sum_{k=1}^{T} p(z = k|x_i, y_l, \Psi^{(t)})\log(\beta_{kl}\theta_{ik})\end{aligned}$$

In the M-step, find the best Ψ that maximizes Q function: $\Psi^{(t+1)} = \arg\max_{\Psi} Q(\Psi|\Psi^{(t)})$, with the constrains on the parameters. By standard calculation with the help of Lagrange multipliers, the updated formulas for Ψ are:

$$\theta_{ik}^{(t+1)} = \frac{\alpha_{i,k}^{0} - 1 + \sum_{j=1}^{N} w_{ij}\theta_{jk}^{(t)} + \sum_{l=1}^{M} c_{il}p(z = k|x_i, y_l, \Psi^{(t)})}{\sum_{k=1}^{T}\alpha_{i,k}^{0} - T + \sum_{j=1}^{N} w_{ij} + \sum_{l=1}^{M} c_{il}} \tag{10}$$

$$\beta_{kl}^{(t+1)} = \frac{\sum_{i=1}^{N} c_{il}p(z = k|x_i, y_l, \Psi^{t})}{\sum_{l'=1}^{M}\sum_{i=1}^{N} c_{il'}p(z = k|x_i, y_{l'}', \Psi^{(t)})} \tag{11}$$

By iteratively applying Eqs. (9), (10) and (11), a local maximum of Ψ will be achieved. By observing Eq. (10), it is interesting to see that at each iteration, θ_i uses two parts of information to update itself, one is from structural information and the other is from text information. What is more, the structural information used in the updating is just the Dirichlet parameter $\boldsymbol{\alpha}_i$ for $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{N_{out}(i)})$, which means at each iteration out neighboring topic distributions are used as priors to derive posterior topic distribution for x_i given the observation of text in the document x_i.

D. Discussions of MRF Modeling on Network Structure

In Section III-B, we have given a new topic model, iTopicModel, that integrates both text information and structural information among documents. We decompose the joint distribution of text and structure into two independent components, one is the structure layer modeled with MRF, and the other is the text layer given the current structure parameter modeled as traditional topic model. For the structure layer, we define an MRF using Eq. (6). Actually, different MRF models can be used in the structure layer of our framework, with different intuitions.

Now we give another MRF definition over the structural layer using both conditional probability and joint probability definition, and then relate it with a newly proposed graph regularization-based topic model method NetPLSA [8].

- Conditional probability definition (local property):

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}) = \frac{\exp\{-\frac{1}{2}\sum_{j\in N(i)} w_{ij}||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||^2\}}{\sum_{\boldsymbol{\theta}_i'}\exp\{-\frac{1}{2}\sum_{j\in N(i)} w_{ij}||\boldsymbol{\theta}_i' - \boldsymbol{\theta}_j||^2\}}$$

  where $||.||$ is the L^2 norm of vector. The intuition of this definition is that the smaller the weighted sum of distance between $\boldsymbol{\theta}_i$ and its neighbors, the higher probability it is, where the distance is evaluated by square of Euclidean distance. Also, the larger strength of a link of two nodes, the closer the two node variables should be.

It can be proved that an equivalent joint distribution can be defined in the following global definition.

- Joint probability definition (global property):

$$p(\boldsymbol{\theta}) = \frac{1}{Z}\exp\{-\sum_{\langle i,j\rangle\in E} V_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}$$

  where $V_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{1}{2}w_{ij}||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||^2 = \frac{1}{2}w_{ij}\sum_{k=1}^{T}(\theta_{ik} - \theta_{jk})^2$ and Z is the partition function $\sum_{\boldsymbol{\theta}}\exp\{\sum_{\langle i,j\rangle\in E} V_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}$.

From the definition, we can see that the direction of links are not considered, and even directed networks will be considered as indirected networks. Under this joint distribution of $\boldsymbol{\theta}$, the log-likelihood can be derived as (with constant normalization part Z neglected):

$$\log L = -\frac{1}{2}\sum_{\langle i,j\rangle\in E} w_{ij}\sum_{k=1}^{T}(\theta_{ik}-\theta_{jk})^2 + \sum_{i=1}^{N}\sum_{l=1}^{M} c_{il}\log(\sum_{k=1}^{T}\beta_{kl}\theta_{ik}) \tag{12}$$

Comparing Eq. (12) with objective function that is to be maximized given by Eq. (6) in NetPLSA [8], which we rewrite using the notations in this paper in the following,

$$\begin{aligned} O(X,G) = &-\frac{\lambda}{2}\sum_{\langle i,j\rangle\in E} w_{ij}\sum_{k=1}^{T}(\theta_{ik}-\theta_{jk})^2 + \\ &(1-\lambda)\sum_{i=1}^{N}\sum_{l=1}^{M} c_{il}\log(\sum_{k=1}^{T}\beta_{kl}\theta_{ik}) \end{aligned} \tag{13}$$

we then can find that this log-likelihood (Eq. (12)) is a special case of Eq. (13) with a fixed parameter $\lambda = \frac{1}{2}$. It is easy to see, we can multiply potential functions $V_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ with constant c to get other λ in Eq. (13). Usually, this constant is represented as $\frac{1}{T}$ in MRF, where T is the temperature of the system. When temperature T is very high, a higher energy state $U(\boldsymbol{\theta})$ is allowed, and in our case a more irregular network structure is allowed. Notice that, a similar parameter can be used in iTopicModel. As we can see, NetPLSA can be integrated into the framework of our model, with a different definition of Markov Random Field, but with the limits that can only deal with undirected networks. Other definitions may also be used under this framework, but need careful reasoning and tests.

IV. Discussions: Practical Issues

In this section, we discuss several practical issues in real applications, such as how to decide the number of topics, how to build a concept hierarchy, and how to choose among several possible networks. Experiments are provided in the later section to demonstrate these issues.

A. Deciding the Number of Topics

How to set the number of topics in topic modeling is always a challenging problem. One possible method is varying the topic number T and choosing the one that maximizes the log-likelihood $\log(L|T)$, such as in Griffiths' work [16]. However, in our model, partition function Z is usually intractable to calculate, and it is a function of T as well. It is difficult to get the exact value of $\log(L|T)$. In network clustering, there is a well known modularity function called Q-function [3], which provides a measure to evaluate the goodness of a clustering on the network. In the original work, they only considered the network with either 0 or 1 weights for edges. Now we extend it to networks with non-negative weights of edges. In our method, topic distribution for each document could be viewed as a soft clustering result. Then we map it to a hard clustering by assigning the cluster label with the highest probability to each x_i. Let $C_{N\times T}$ denote the hard clustering indicator matrix, we thus give the Q-function over network G given $\boldsymbol{\theta}$ as:

$$\begin{aligned} Q(C|G) &= \sum_{k=1}^{T}\Big(\frac{w(X_k,X_k)}{D} - (\frac{w(X_k,X)}{D})^2\Big) \\ &= \sum_{k=1}^{T}\Big(\frac{C(:,k)'WC(:,k)}{D} - (\frac{\sum_{j=1}^{T}C(:,k)'WC(:,j)}{D})^2\Big) \end{aligned} \tag{14}$$

where X_k denotes the set of nodes belonging to cluster k, $w(X_k, X_k)$ denotes the total weights of links whose both nodes are in cluster k, $w(X_k, X)$ denotes the total weights of links that contains at least one node in cluster k, and $D = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$ is the total weights of all the edges. This formula calculates the difference between within-cluster percentage of edges and the percentage of edges in a random case. Q varies from 0, when the network is totally random and has no clustering structure; to approaching 1 (actually $1 - 1/T$), when there are no inter-cluster edges at all. According to [3], Q lies in the range of $[0.3, 0.7]$ for networks with strong community structure. By varying the number of topics, we can select the T that maximizes Q-function defined in Eq. (14).

B. Building Concept Hierarchies

Concept hierarchies would be very useful for data analysis. For example, concept hierarchies such as ACM Classification System[6], would help authors to classify their papers, and help users index and search papers in a large collection of bibliographic data. However, it requires a lot of human labor to build such concept hierarchies manually, when the dataset is large. Also, the concept hierarchies may be changing along with time, *e.g.*, ACM Classification System has been changed several times since its first appearance. Therefore, automatically building concept hierarchies that are described using topics, would be very useful for data analysis tasks, *e.g.*, OLAP service.

A key problem in building concept hierarchy using topic modeling in a heterogeneous network that contains multiple types of objects is that, we should carefully select different object types to be the documents in different scales. For example, a conference network should be enough to find the first level topics. While for finer level topics, conference will be too coarse. In contrast, it may not be wise to use papers to get the first level topics, since they contain too detailed information and are not able to generate the overall view of the data. We will illustrate how to build a concept hierarchy using the DBLP data as an example in the experiment part. It turns out that, by recursively applying iTopicModel on the sub document network, with Q-function as the branch selection measure, a concept hierarchy can be automatically built.

[6] http://www.acm.org/about/class/

C. Effects of Network Structures

For the same document corpus, there are multiple ways to construct networks among documents. For a collection of papers in bibliographic database, paper citation network introduced in Ex. 3.1 can be built. Similarly, networks can also be constructed based on co-author relationships, or through text similarity. Consider the following two extreme cases of the relation between network and text information:

1) The links of the network among documents are randomly formed, and in this case network structure will not help topic modeling, and even deteriorate the performance of results.
2) The links of the network among documents are built exactly through the text information, and in this case network structure will not improve the topic modeling performance too much.

In order to measure how close a network over the documents to the text information this document corpus contains, we propose a correlation measure between network and text information:

$$Corr(G, X) = \frac{\sum_{\langle x_i, x_j \rangle \in E} w_{ij} sim(x_i, x_j)}{\sqrt{\sum_{\langle x_i, x_j \rangle \in E} w_{ij}^2} \sqrt{\sum_{\langle x_i, x_j \rangle \in E} sim(x_i, x_j)^2}} \quad (15)$$

where G, X, x_i, x_j are defined as in Section III-A. $sim(x_i, x_j)$ is the similarity measure for two documents, and it is defined as $\cos(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ is the word count vector defined in Section III-A. We can also use *tfidf* instead of word count in the representation vector. How correlation between network and text impacts the topic modeling performance will be studied in Section V-D.

Another question would be, shall we trust all the links in the network? The answer will be no. According to our experimental study, a network that has too many small weight links will degrade the performance of iTopicModel, since small weight links may happen occasionally, only link with a larger weight shows a consistent, strong relation among documents. So we transform our original networks into a KNN network, which means a document only keeps K most connected documents as neighbors, according to the weight of links.

V. Experiments

In this section, we apply our model to several real datasets, and show its usefulness in real applications and its effectiveness over several state-of-the-art models. Also, we study the impacts of different network structures on the topic modeling, which could be served as a hint to choose network structures in real cases.

A. Datasets

We use two datasets in the experiments, the DBLP dataset and the Cora Research Paper Classification dataset[7]. For the DBLP data, we extract two datasets: (i) the "all-area" data set, which contains top 1000 conferences and top 50000 authors by their publication numbers, and all the publications of these authors; and (ii) "four-area" dataset, which includes 20 major conferences from four related areas, *i.e.*, database, data mining, machine learning and information retrieval, and all the 28702 authors and their publications in these conferences. For the Cora dataset, after preprocessing, we get 19396 papers with their citation lists, author lists and title information. Each paper in Cora has a classification label from total 70 classes. Notice that, for both datasets, we only have titles as text information for papers.

[7]http://www.cs.umass.edu/∼mccallum/code-data.html

B. Building Concept Hierarchies in DBLP

In this case study, we use the DBLP "all-area" dataset to build concept hierarchies in Computer Science. Different document networks can be obtained from the data. For conferences, network is derived from shared author numbers between conferences, and document for each conference is the compacted titles of all the papers in that conference. For authors, networks is formed by the co-authorship, *i.e.*, the number of papers they co-authored, and the document for each author is the grouped titles of all the papers that author published. For papers, networks can be formed by either the similarity of text between paper titles or the shared author number between papers, and the title itself serves as a document. As discussed in Section IV-B, for the first level, we use conference network to generate the most coarse topics. By using the modularity measure Q-function defined as in Eq. (14), we found 7 is the best topic number. The Q-function measure varying with different topic number T is shown in Figure 3. The seven topics are summarized in Table II by top-10 words in each topic. For further modeling on a parent topic, we use conferences that have max probabilities in that topic as a constraint to select top 1000 authors appearing in those conferences and build the author network. The subtopics in the subarea of Topic 4 at the first level is shown in Table III.

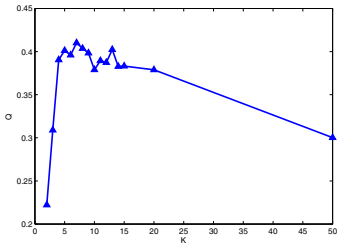

Figure 3. Q-function vs. Topic Number T

C. Performance Study on Topic Modeling

How to judge whether a topic modeling is good? One typical method is to print the top-ranked words and judge

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
network	network	parallel	system	system	graph	image
system	model	design	data	software	problem	recognition
performance	algorithm	test	model	design	algorithm	base
distribute	neural	simulation	database	web	set	model
wireless	fuzzy	high	language	information	bound	3d
protocol	method	architecture	base	computer	tree	video
control	analysis	power	program	model	complexity	detection
scheme	learn	circuit	approach	environment	linear	robot
channel	problem	memory	query	management	order	motion
mobile	genetic	fault	knowledge	development	function	segmentation

Table II
FIRST LEVEL TOPICS IN DBLP (BEST T = 7, TOP-10 WORDS FOR EACH TOPIC)

Topic 4.1	Topic 4.2	Topic 4.3	Topic 4.4	Topic 4.5	Topic 4.6
logic	web	data	learn	database	agent
program	information	mine	algorithm	data	system
reason	retrieval	efficient	network	system	plan
constraint	model	cluster	model	query	model
knowledge	semantic	query	data	object	learn
set	system	index	analysis	xml	multiagent
semantics	language	algorithm	structure	management	knowledge
theory	document	network	protein	model	problem
model	knowledge	search	tree	process	approach
language	base	database	sequence	distribute	robot

Table III
SUB TOPICS OF TOPIC 4 IN LEVEL 1 (BEST T = 6, TOP-10 WORDS FOR EACH TOPIC)

them by experience. We use two measures to compare the results of different topic modeling methods: (i) NMI (Normalized Mutual Information) [17] that measures the goodness of document clustering results mapped from topic distribution over documents $\boldsymbol{\theta}$ by comparing them with the human-labeled clustering results, and (ii) Q-function measure as defined in Eq. (14) that measures the consistency of the clustering results over the network. NMI is used to compare two clustering results, say $\mathcal{C}$ and $\mathcal{C}'$, without knowing the mapping relation between them:

$$NMI(\mathcal{C},\mathcal{C}') = \frac{\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'} p(i,j)\log(\frac{p(i,j)}{p_{\mathcal{C}}(i)p_{\mathcal{C}'}(j)})}{\sqrt{\sum_{i\in\mathcal{C}} p_{\mathcal{C}}(i)\log p_{\mathcal{C}'}(j)\times\sum_{j\in\mathcal{C}'} p_{\mathcal{C}'}(j)\log p_{\mathcal{C}'}(j)}} \tag{16}$$

where i and j are cluster labels in $\mathcal{C}$ and $\mathcal{C}'$, $p(i,j)$ is the percentage of shared common objects in both clusters i and j, $p_{\mathcal{C}}(i)$ the percentage of objects in i in clustering $\mathcal{C}$, and $p_{\mathcal{C}'}(j)$ the percentage of objects in j in clustering $\mathcal{C}'$. NMI is in the range of $[0,1]$, and higher value means higher agreement among two clusterings.

We use three different networks, conference net and author net from DBLP, and paper citation net from Cora. Conference net and author net (top 1000 authors used) are extracted from the DBLP "four-area" dataset using the methods described in Section V-B, named as "ConfNet" and "AuthorNet" respectively. For the paper citation net, we selected five classes in the level of whole computer science, named as "PaperNet-Cite". We labeled the 20 conferences and 200 authors sampled from the 1000 authors to the four areas for "ConfNet" and "AuthorNet", and use the classification labels for papers from Cora data.

Four topic modeling methods, PLSA, LDA, netPLSA, and iTopicModel, are studied. We use topic modeling toolbox [18] for the LDA method. The results are summarized in Tables IV and V. All the results are based on 10 rounds running of each algorithm. The experiments show that, in the measure of NMI, iTopicModel outperforms or has comparable performance in all the datasets, and especially good at the document network that with very short text information in each document, such as "PaperNet". We can see that, without the information of links among papers, PLSA and LDA can rarely get the right clusters for each document in "PaperNet". But for NetPLSA and iTopicModel, since they have considered the network structure, the clustering results are much better; and the latter is even better than the former in most datasets, since it has a better MRF definition to model the dependency relation among documents. Notice that, the NMI is not that high partly because the classification data provided in Cora is not that accurate, warned by the data provider. Also, we actually can pick the best result among several runnings according to the final $logL$ (the larger the better). In our case for dataset "PaperNet-Cite", we can get results with NMI above 0.5. In the measure of Q-function, iTopicModel consistently provides better topics that are consistent with the network structure.

D. Network Structure Study

We build three different networks for papers using the dataset of Cora, and study the relationship between their correlations to the text and the topic modeling performance. Besides the citation networks, we also construct networks for papers using the co-author number, *i.e.*, "PaperNet-Author",

	PLSA	LDA	NetPLSA	iTopicModel
ConfNet	0.7959	0.7469	0.7291	**0.8255**
AuthorNet	0.4059	**0.5639**	0.4761	0.5360
PaperNet-Cite	0.1287	0.0674	0.4291	**0.4424**

Table IV
DOCUMENT CLUSTERING ACCURACY: NMI

	PLSA	LDA	NetPLSA	iTopicModel
ConfNet	0.4364	0.4370	0.4368	**0.4440**
AuthorNet	0.4059	0.4910	0.4760	**0.4938**
PaperNet-Cite	0.1703	0.0984	0.4760	**0.5783**

Table V
CLUSTERING CONSISTENCY: Q-FUNCTION

and co-text number, *i.e.*, "PaperNet-Text'. Co-text number means the co-occurrence word number for any two papers. The results are summarized in Table VI. From the results we can see that, for paper network built from text, the clustering performance is very similar to topic modeling using only text information (see NMI result in Table IV for PLSA). Actually, "ConfNet" also has a high correlation between network and text, and thus the performance for iTopicModel has not improved too much from PLSA for this network.

	PaperNet-Cite	PaperNet-Author	PaperNet-Text
NMI	0.4424	0.2329	0.1404
Q	0.5783	0.6310	0.3253
Corr	0.1719	0.1420	0.7658

Table VI
NETWORK STRUCTURE STUDY

VI. CONCLUSIONS

In this paper, a new framework of generative topic model called iTopicModel is proposed that integrates both network structure and text information for the popularly encountered document networks. This model has a two-layer graphical model structure. On the top, a reasonable multivariate Markov Random Field is defined to model the dependency relations among documents. On the bottom, a traditional document generative model is used, which is conditionally independent with each other given the current topic distribution configurations. A joint probability function is then defined based on the graphical model. We then propose an EM-based iterative solution to estimate the set of best parameters that maximizes the log-likelihood of the joint distribution. Our experiments show that this model is more effective than the state-of-the-art topic modeling methods, in both aspects: following human intuition and being consistent with the network structure. Also, we show that this model, with the help of Q-function, can help us automatically build concept hierarchy in online databases. The future work could be on how to build topic models that integrate different network structures with text information.

REFERENCES

[1] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI'99*, Stockholm, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[3] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 69, no. 2, 2004.

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[5] U. von Luxburg, "A tutorial on spectral clustering," Max Planck Institute for Biological Cybernetics, Tech. Rep., 2006.

[6] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *AUAI'04*. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494.

[7] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *KDD'04*. New York, NY, USA: ACM, 2004, pp. 306–315.

[8] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *WWW'08*. New York, NY, USA: ACM, 2008, pp. 101–110.

[9] J. Chang and D. M. Blei, "Relational topic models for document networks," in *AISTATS'09*, April 2009, pp. 81–88.

[10] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in *CIKM'06*, 2006, pp. 778–779.

[11] R. Angelova and G. Weikum, "Graph-based text classification: learn from your neighbors," in *SIGIR'06*, 2006, pp. 485–492.

[12] P. Perez, "Markov random fields and images," *CWI Quarterly*, vol. 11(4), pp. 414–437.

[13] R. Kindermann, *Markov Random Fields and Their Applications (Contemporary Mathematics; V. 1)*. American Mathematical Society, 1980.

[14] S. Z. Li, *Markov random field modeling in image analysis*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[15] T. P. Minka, "Estimating a dirichlet distribution," 2003. [Online]. Available: http://research.microsoft.com/˜minka

[16] T. L. Griffiths and M. Steyvers, "Finding scientific topics." *Proc Natl Acad Sci USA*, vol. 101 Suppl 1, pp. 5228–5235, April 2004.

[17] A. Strehl, J. Ghosh, and C. Cardie, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[18] M. Steyvers and T. Griffiths, "Matlab topic modeling toolbox." [Online]. Available: http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm