Mining Topics on Participations for Community Discovery

Guoqing Zheng[†], Jinwen Guo[†], Lichun Yang[†], Shengliang Xu[†],
Shenghua Bao[‡], Zhong Su[‡], Dingyi Han[†], Yong Yu[†]

[†]Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, 200240, China
{gqzheng, guojw, lichunyang, slxu, handy, yyu}@apex.sjtu.edu.cn

[‡]IBM China Research Laboratory
Beijing, 100094, China
{baoshhua, suzhong}@cn.ibm.com

ABSTRACT

Community discovery on large-scale linked document corpora has been a hot research topic for decades. There are two types of links. The first one, which we call *d2d-link*, indicates connectiveness among different documents, such as blog references and research paper citations. The other one, which we call *u2u-link*, represents co-occurrences or simultaneous **participations** of different users in one document and typically each document from u2u-link corpus has more than one user/author. Examples of u2u-link data covers email archives and research paper co-authorship networks. Community discovery in d2d-link data has achieved much success, while methods for that in u2u-link data either make no use of the textual content of the documents or make oversimplified assumptions about the users and the textual content. In this paper we propose a general approach of community discovery for u2u-link data, i.e., multiple user data, by placing topical variables on multiple authors' participations in documents. Experiments on a research proceeding co-authorship corpus and a New York Times news corpus show the effectiveness of our model.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms, Experimentation

Keywords

Topics on Participations, Community discovery, Nonparametric statistical model, Hierarchical Dirichlet Process

SIGIR'11, July 24–28, 2011, Beijing, China.

1. INTRODUCTION

When we talk about linked document data, there are two different types of links. Link of the first one, which we call *d2d-link*, indicates connectiveness between different documents. These kind of data include blog reference data and research paper citation data. Link of the other type, which we call *u2u-link*, represents co-occurrence or simultaneous participations of different users in one document. Usually there are more than one user in each document of this data, so we also call it *Multiple User Data*(MUD). Email archives and research paper co-authorship network fall into this category. Figure 1 gives a sample of a research proceeding corpus, where $u_1, \cdots, u_4$ denote four researchers, and d_1, d_2, d_3 denote three research papers with different colors indicating different research areas. Many methods of community discovery in d2d-link data have been proposed and gain satisfactory results. But as to u2u-link data, rare solutions which combine link information and textual content for community discovery exist. Current approaches either make no use of the textual content of the documents or make oversimplified assumptions about the users as well as the textual content.

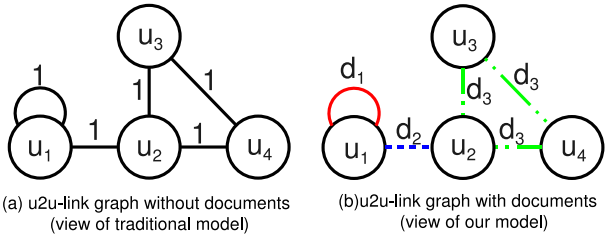

Figure 1: Different views of u2u-link graph(best viewed in color)

One part of traditional methods of community discovery in u2u-link data are performed solely on a u2u-link graph in which the vertices represent the users and the edges indicate links between pairs of authors. A community is typically defined as a subset of users with better connectivity amongst its members than between its members and other users of the graph. The task of community discovery is then to recover such subsets from the given graph. Most of these approaches, including graph cut based methods[21], modularity based methods[16], flow based methods[6] and spectral based methods[20, 1], typically choose an objective function which captures the above intuition of a community and then try to optimize the objective function[9]. These methods only model the links by assigning a certain weight

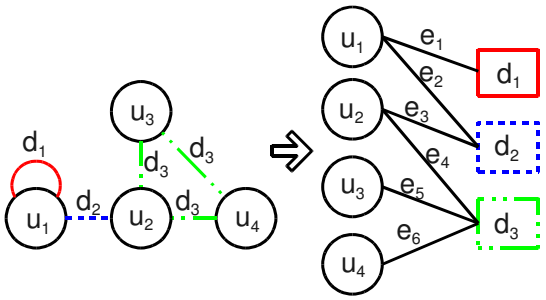

Figure 2: A u2u-link graph and its corresponding participation graph(best viewed in color)

to each link between pairs of users, as the co-authorship network shown in Figure 1(a); though widely applicable and accepted, they make no use of the textual content of the documents and thus give no clue how different authors participate in the documents and what semantics these participations imply. Though some other methods take the textual content into account, they make oversimplified assumptions and thus ignore useful participation information. The NetPLSA model[15] constructs the u2u-link graph as described in Figure 1(a), merges all documents one user participates in into a single document for that user. Thus NetPLSA ignores the various participation information for each user. The Author-Topic(AT) model[22] learns the topic of a document conditioned on the mixture of authors(users) of that document, which implicitly assumes that the distribution over topics of a certain author will not change for all his/her documents. This assumption seems to suffice but it may not fit some real cases such as a researcher switches to another research area and publishes research papers with different topics. A better modeling may exists.

In particular, we propose a Topics-on-Participations(ToP) model for community discovery in u2u-link data. First, we transform the u2u-link graph to an equivalent *participation graph* for clarity, as illustrated in Figure 2. To better exploit the multi-author attribute of each document in the data, we place topical variables on each participation each author get involved in. More specifically, given the u2u-link data, ToP leverages Hierarchical Dirichlet Process(HDP) to introduce hidden topical variables of the participations. Other than the use of participation structure, ToP can look into the documents of the participations in detail to recover sound communities. We first generate the hidden topic variables of the participations and then documents are generated by these according to the hidden variables. Besides, ToP is a nonparametric model, which is a natural outcome of HDP. The setting of community number is a common problem for most of the community discovery algorithms proposed so far. Due to the nonparametric property, ToP is able to automatically select the proper community number according to the observed data.

We evaluate ToP on two real-world datasets, a research proceeding corpus and a New York Times news corpus. The research proceeding corpus consists of paper abstracts from 7 research conferences, i.e. ACL, ICML, SIGGRAPH, SIGIR, SIGKDD, SIGMOD, and WWW during 2005-2009. In total there are 9415 individual authors and 5308 documents. And the New York Times news corpus contains business news mentioning selected sets of companies. As the whole corpus is large, we collect a subset of the corpus, which contains 1677 companies and 2461 documents. Quantitative evaluation on both two datasets shows that ToP outperforms two related community discovery algorithms. Deeper analysis shows that ToP can not only identify the proper number of communities, but also reveal reasonable semantic interpretation for every discovered community.

The main contribution of this paper is two-fold:

- The proposal of ToP which models topics on participations for community discovery in u2u-link data. This modeling makes it possible to mine detailed community information from participations.
- The extensive evaluation on two real-world datasets, verifies the effectiveness of the proposed model. It can automatically detect the proper number of communities, provide a reasonable interpretation of the discovered communities.

The remainder of this paper is organized as follows: Section 2 introduces some related work. Section 3 gives some definitions and formally define the community discovery task. Section 4 presents ToP model in detail. Experimental results as well as some case studies are presented in Section 5. At last, we conclude the paper and discuss about future work in Section 6.

2. RELATED WORK

Much work has been done on community discovery. In this section, we review two lines of work: community discovery in d2d-link data and community discovery in u2u-link data.

2.1 Community Discovery in d2d-link data

In the context of d2d-link data, probabilistic models, such as Latent Dirichlet Allocation(LDA)[3] and its variations, play an important role in uncovering underlying topics from textual data. Zhou et.al [26] successfully extract e-communities based solely on the content of communication documents. Their work gives us a strong support in mining text contents for community discovery. With the aim of taking into account the link information among documents, serveral methods have been proposed. Li et al. in [10] design a sophisticated model for scalable community discovery on textual data with relations. Their model explores the d2d-link graph to detect some community cores and then uses text information to improve community consistency. Cohn and Hofmann combine PLSA and PHITS together and derive a unified model from text contents and citation information of documents under the same latent space [4]. Erosheva et al. apply similar mixed membership model [5] for soft clustering of papers by using text content and references. Liu et al. in [12] propose the Topic-Link LDA model with the intuitive that a link between two documents is determined by both content similarity and author community membership similarity and gain improved results against previous methods. All these approaches captures several aspects of the d2d-link data, but they may be not suitable for community discovery in the u2u-link data.

2.2 Community Discovery in u2u-link data

In the past few decades, many models have been proposed for community discovery in the u2u-link data, but most of

these methods are based solely on the u2u-link graph and make no use of the textual content of the documents. Note here, as these methods take into account only the link information and do not consider textual content , they can also be applied to the above mentioned d2d-link data, depending on how a "user" is defined.

Traditionally, community discovery is performed by partition of an u2u-link graph with the aim of optimizing a selected objective function. Many methods have been proposed along this direction, e.g., graph cut based methods, flow based methods, modularity based methods, spectral clustering based direction, etc. Graph cut based methods, including NCut[21], try to find an optimal graph partition with the edge weight between partitions minimized or edge weight inside a partition maximized. Due to the NP-complete complexity of this method, approximate solutions have been proposed. Flake et al.[6] propose approximate algorithms based on network flow ideas to partition the network by solving maximum flow problems, where they define community as a set of entities that has small inter-community cuts and large intra-community cuts. Girvan and Newman[7] introduce *betweenness centrality* to detect communities. After that, Newman and Girvan[16] introduce *modularity* to measure the overall quality of discovered communities. Modularity evaluates how entities in a community connect with other entities in that community and has been adopted by many community detection literature. Modularity can be optimized by using the eigenvectors of the modularity matrix which gives rise to those spectral clustering based methods[20, 1]. McCallum et al. in [14] study a new community discovery task on u2u-link data. Instead of finding densely connected entities, they seek to find out users with similar connection pattern such as similar voting patterns.

Another part of previous work analyzing u2u-link data is based on statistical models[18, 8, 11]. In these models, a latent variable is introduced for each entity to express its potential characteristic and then the models generate the links between them according to the latent variable assignments which are most likely to recover the graph structure[18]. Kemp et al. in [8] extend these models to infinite latent classes. The model can automatically determine the proper latent class number based on the observed graph structure. However, it restricts each entity to only one latent class. Additionally, the discovered class/group in these models is a set of entities which have similar connectivity pattern in the graph, which is quite different from the traditional definition of community, i.e., a densely connected subgraph of the whole network. Because of their general modeling property, these algorithms have been widely adopted, such as biology data analysis [24, 19], co-authorship network analysis [17], Web community identification [6]. Although there are many other great studies in this direction, we do not give a detailed review for each of them because they do not utilize the text information to enhance the community discovery.

In the context of u2u-link data, there have also been some related studies aiming to take advantages of both text contents and the u2u-link structure. Most of these algorithms perform community discovery on a u2u-link graph with text user profiles. In [15], Mei et al. propose a regularized model NetPLSA for discovering more smoothing topics/communities over u2u-link graph by constraining that connected documents that have similar topic distributions. The Author-Topic(AT) model uncovers topics conditioned on the mixture of authors that composed a document[22]. McCallum et al. modeled email archives by a generative process [13]. The proposed Author-Recipient-Topic(ART) model gives each pair of author and recipient a topic distribution instead of a topic distribution per-author. Though their modeling setting shares some similarity with ours, their model is specifically designed for email archives. Different from the above models, our method tries to handle the setting where the text contents are usually associated with more than one user and we aim to quantify the effect of each user's participation contribution to the documents' topical distribution.

3. PROBLEM STATEMENT

In this section, we formally give some definitions and define the task of community discovery from u2u-link data.

Definition 1. **(Participation Graph)**: A participation graph $G = (U \cup \mathcal{D}, E)$ is a bipartite graph where the set of users U and the set of documents $\mathcal{D}$ are two disjoint sets. An edge $e \in E$ connecting user $u \in U$ and document $d \in \mathcal{D}$ indicates that u participates d and we denote this participation with e. So we can construct an equivalent participation graph from every u2u-link data. As we see in Figure 2, we transform a participation graph from the sample u2u-link graph.

Traditionally, a community is a group of users with dense interactions amongst its members than between its members and the remainder of the u2u-link graph. While in our setting, since we have two information sources for community identification, we need to expand the community definition to incorporate both of them, i.e. the topical community.

Definition 2. **(Topical Community)**: A topical community is a soft partition of the users with a multinomial word distribution over the vocabulary $\mathcal{V}$. We denote the topical community space as ϕ_∞. The soft partition means that each user is assigned to each of the communities with a membership weight, i.e. a community distribution. This modeling is natural since it is always the case that a user can belong to multiple communities.

Definition 3. **Task (Community Discovery)**: Given an participation graph G , the task of *Community Discovery* is to find a set of topical communities $\{c_1, c_2, ..., c_K\}$, where the community number K is detected automatically and to calculate the community distribution of each user u, i.e. $\boldsymbol{\theta_u}$.

4. COMMUNITY DISCOVERY MODELING

Briefly speaking, our model looks into the participations, place topical variables on participations and analyzes the latent topics of the documents involved in those participations. In this section, we will introduce our model, the inference method and some basic analysis as well.

4.1 Topics on Participations

Our Topics on Participations(ToP) model first utilizes Hierarchical Dirichlet Process to introduce latent topic variables to each user participation of a document, which forms the modeling for the participation graph part. Then, for the document part, we specify the generation process of a document via a simple topic modeling technique. Finally, we

combine the two separate modeling by assuming that they share the same latent topic space.

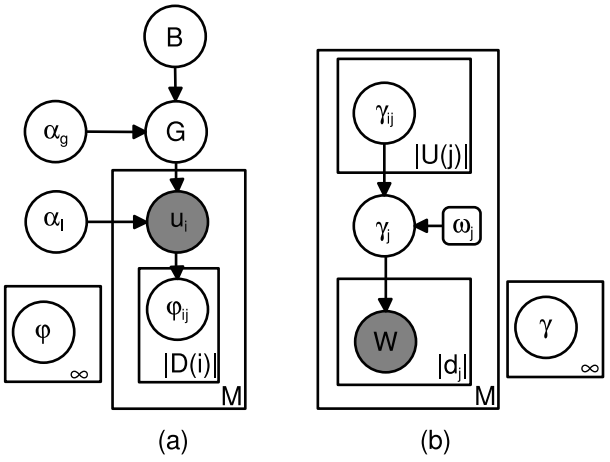

Figure 3: Separated graphical model of ToP. (a) models the participation graph, (b) models the documents.

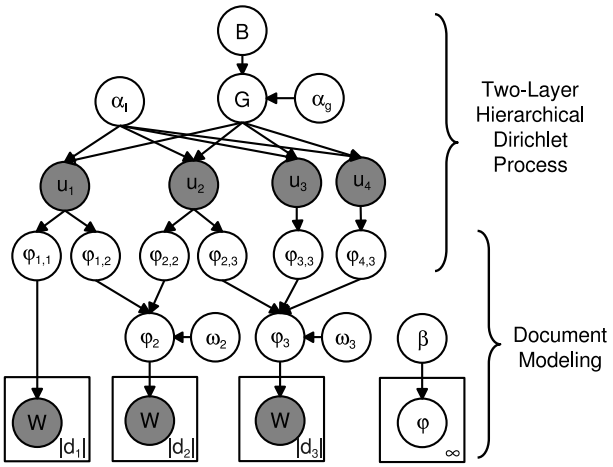

Figure 4: Combined graphical model of ToP for the sample graph in Figure 2 (b)

Participation Graph Modeling. From the participation graph view, we may consider that each user u's participation $e_{u,d}$ in a document d is an evidence indicating the community for the user. In our model, we choose to make it explicit by setting a topical variable for each participation. We call these variables *evidence variables*. Since the community number is typically not known beforehand, $e_{u,d}$ should be a draw from an infinite semantic community space, which will allow the model to learn the community number through the data. Here, we utilize Hierarchical Dirichlet Process to assign evidence variables and build our model.

Figure 3(a) shows the graphical model of our Participation Graph Modeling, where M is the number of users and $|D(i)|$ is the number of documents linked to user i. The outside global Dirichlet Process provides a shared infinite number of variables for all the users' evidence variables. This infinite space brought the non-parametric manner, which enables the model to learn topic number according to the real data. Formally, the Dirichlet process is

$$G|\alpha_g, B \sim DP(\alpha_g, B) \tag{1}$$

where B is the *base measure* and α_g is a positive real number called the *concentration parameter*.

The inside local Dirichlet Process models all participation of a single user in his/her documents. This models the clustering property of a user's participation in documents, i.e. if most publications of a user are related to information retrieval, the next paper of the user will most probably to be about information retrieval again. Formally,

$$U|\alpha_l, G \sim DP(\alpha_l, G) \tag{2}$$

where α_l is a positive real number, called *concentration parameter*. Then, every user's evidence variables are drawn from the infinite space ϕ_∞ as follows:

$$p(\phi_{i,j_{n+1}}|\phi_{i,j_1}, \cdots, \phi_{i,j_n}; \alpha_l) = \frac{\alpha_l G + \sum_{h=1}^{n} \delta(\phi_{i,j_h})}{n + \alpha_l} \tag{3}$$

where ϕ_{i,j_h} denotes the hth evidence variable of U_i, and δ is the atom function. This actually forms a two-layer Hierarchical Dirichlet Process. The specific modeling details and reason of modeling choices can be found in [23].

Document Modeling. To explore the semantics of the text documents, topic modeling is a good choice like PLSA and LDA. Similarly, we set a hidden document model for each document and assume that all the words in that document are drawn from the hidden model. Moreover, since each document is generated by one author or a group of authors, we add a layer that chooses a certain author's topic.

Figure 3(b) shows the graphical model of our Document Modeling, where $|d_j|$ is the number of words in d_j and $|U(j)|$ denotes the number of topic variables connected to γ_j. On the top, γs are the topic variables of all the authors of the document, which are draws from the infinite space γ_∞. Then, ω is the weighting vector parameters for the topic selection process, satisfying $\sum_{h=1}^{|U(j)|} \omega_{j,h} = 1$. The relation of these variables is as

$$p(\gamma_j|\gamma_{(1),j}, \gamma_{(2),j}, \cdots, \gamma_{(|U_j|),j}) = \sum_{h=1}^{|U_j|} \omega_h \delta(\gamma_{(h),j}) \tag{4}$$

where $\gamma_{(h),j}$ denotes the hth evidence variable that connects γ_j. This setting captures the interaction between the authors of the same document by the probability feature that given γ_j=ψ, the more $\gamma_{(h),j}$ having value ψ, the larger the probability of selecting ψ by γ_j. In other words, the topic variables linked to γ_j tend to agree with each other. Then, we assume that all the words in that document are drawn from the topic model, as the lower part of the Figure 3(b). The joint probability of a document model γ_j and its generated words $\boldsymbol{W}_j$ is:

$$p(W_{j,1}, W_{j,2}, \cdots, W_{j,|\boldsymbol{W}_j|}, \gamma_j) = p(\gamma_j) \prod_{h=1}^{|\boldsymbol{W}_j|} p(W_{j,h}|\gamma_j) \tag{5}$$

where γ_j is a draw from an infinite semantic space γ_∞.

Note that, since the modeling focus is to recover the topic of the links between users, we choose this simpler method (comparing to LDA and PLSA) to model the documents. In fact, this method has also been widely used in language model for information retrieval and naive bayes classifiers because of its light weight yet effective.

Combination of Participation Graph Modeling and Document Modeling. Considering that the community space ϕ_∞ and the document model space γ_∞ are both semantic spaces, we simply unite them into a single space, ϕ_∞. Now, the whole modeling process is finished. To make it clearer, the whole model is shown in Figure 4. Note that the model can not be simplified into box representation and we have to expand the variables. Figure 4 just gives a sample according to the participation graph in Figure 2(b).

4.2 Model Inference

Table 1: the symbols which will be used in model inference

M	the number of users
N	the number of documents
$\boldsymbol{W}_j$	the jth observed document
$D_s(i)$	the set of single-user documents linked to user i
$D_m(i)$	the set of multi-user documents linked to user i
$U(j)$	the set of users linked to document j
$T(i)$	the set of tables in restaurant i
$t_{i,j}$	table index of customer (i,j), $i \in U(j)$, $j \in D(i)$
$k_{t_{i,j}}$	the dish serving on the table that $t_{i,j}$ refers
$k_{i,t}$	$t \in T(i)$, the dish on the tth table in restaurant i
d_j	the mixture component (i.e. dish) index of multi-user document j
$m_{i,k}$	number of tables in restaurant i serving dish k
$n_{i,t,k}$	the number of customers in restaurant i, sitting at table t, eating dish k
$var_\cdot$	a marginal of the variable var, e.g. $m_{i,\cdot}$ denotes the number of tables in restaurant i.
$\boldsymbol{var}$	the set of all var, e.g $\boldsymbol{t}$ denotes the table indices of all the customers
var^{-s}	all the var set excluding the upper scripted one, e.g. $\boldsymbol{t}^{-ij}$ denotes all the table indices but t_{ij}

We employ Gibbs sampling for model inference [2]. Firstly, we derive two likelihood expressions for the convenience in the inference derivation later. The conditional density of $\boldsymbol{W}_j$ under mixture component k given all other observed documents is:

$$f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) = \frac{\int Mul(\boldsymbol{W}_j|\phi_k)\prod_{j'\neq j,\left\{\substack{d_{j'}=k,j'\in D_m(\cdot)\\ k_{t_{\cdot,j'}}=k,j'\in D_s(\cdot)}\right\}} Mul(\boldsymbol{W}_{j'}|\phi_k)Dir(\phi_k)d\phi_k}{\int \prod_{j'\neq j,\left\{\substack{d_{j'}=k,j'\in D_m(\cdot)\\ k_{t_{\cdot,j'}}=k,j'\in D_s(\cdot)}\right\}} Mul(\boldsymbol{W}_{j'}|\phi_k)Dir(\phi_k)d\phi_k} \tag{6}$$

where $Mul()$ denotes multinomial distribution. And the selection probability distribution:

$$s(d_j|k_{t_{\cdot,j}}) = \sum_{i'\in U(j)} \delta(k_{t_{i',j}}) \tag{7}$$

We have set all ω's to be uniform so that these parameters can be omitted. Further note that we will suppress references to the variables in the condition part of a conditional probability except those will used in the probability expression.

There are three sets of hidden variables that need to be sampled: table indices $\boldsymbol{t}$ of customers, mixture component indices $\boldsymbol{d}$ of documents, and dish indices $\boldsymbol{k}$ of tables. We derive the Gibbs sampling expression for each of them.

Sampling $\boldsymbol{t}$. The variable set $\boldsymbol{t}$ should be split to two sets because they are different in sampling. Firstly, if the document j is a single user document, $t_{i,j}$ is sampled by combing the likelihood of generating the observed documents.

$$p(t_{i,j}=t|\boldsymbol{t}^{-i,j},\boldsymbol{k}) \propto \begin{cases} n_{it\cdot} f_{k_{i,t}}^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) & if\ t\ is\ previously\ used \\ \alpha_l p(\boldsymbol{W}_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) & if\ t=t^{new} \end{cases} \tag{8}$$

where $p(\boldsymbol{W}_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k})$ is the liklihood for $t_{i,j}=t^{new}$:

$$p(\boldsymbol{W}_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) = \sum_{k=1}^{K}\frac{m_{\cdot,k}}{m_{\cdot,\cdot}+\alpha_g} f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) + \frac{\alpha_g}{m_{\cdot,\cdot}+\alpha_g} f_{k^{new}}^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) \tag{9}$$

If the sampled value of $t_{i,j}$ is t^{new}, we need to obtain a sample of $k_{i,t^{new}}$:

$$p(k_{i,t^{new}}|\boldsymbol{t},\boldsymbol{k}^{-i,t^{new}}) \propto \begin{cases} m_{\cdot,k} f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) & if\ k\ is\ previously\ used \\ \alpha_g f_{k_{new}}^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) & if\ k=k^{new} \end{cases} \tag{10}$$

Secondly, for $\boldsymbol{t}$ of multi-user documents, we have a similar sampling process but the likelihood function is replaced by the selection function in Equation 7.

$$p(t_{i,j}=t|\boldsymbol{t}^{-ij},\boldsymbol{k},\boldsymbol{d}) \propto \begin{cases} n_{i,t,\cdot}^{-i,j} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k_{i,t}) & if\ t\ is\ previously\ used \\ \alpha_l p(d_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) & if\ t=t^{new} \end{cases} \tag{11}$$

where $p(d_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k})$ is:

$$p(d_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) = \sum_{k=1}^{K}\frac{m_{\cdot,k}}{m_{\cdot,\cdot}+\alpha_g} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k) + \frac{\alpha_g}{m_{\cdot,\cdot}+\alpha_g} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k^{new}) \tag{12}$$

And in the case of choosing t^{new}:

$$p(k_{i,t^{new}}|\boldsymbol{t},\boldsymbol{k}^{-i,t^{new}}) \propto \begin{cases} m_{\cdot,k} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k) & if\ k\ is\ previously\ used \\ \alpha_g s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k^{new}) & if\ k=k^{new} \end{cases} \tag{13}$$

Sampling $\boldsymbol{d}$. These variables relate only to the selection processes and the multi-user document likelihood. They are thus sampled as

$$p(d_j=k|\boldsymbol{W}_\cdot,\boldsymbol{d}^{-j},\boldsymbol{t},\boldsymbol{k}) \propto s(d_j=k|k_{t_{\cdot,j}}) f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) \tag{14}$$

Sampling $\boldsymbol{k}$. Since changing $k_{i,t}$ will change the mixture components of all the $t_{i,\cdot}$, the sampling relates to both the selection likelihood and the document likelihood of these

variables. So the Gibbs sampling expression is:

$$p(k_{i,t} = k|\boldsymbol{t}, \boldsymbol{k}^{-it}) \tag{15}$$
$$\propto \begin{cases} m_{\cdot,k}^{-it} f_k^{-\boldsymbol{W}_{\{j,j\in D_s(i)\}}}(\boldsymbol{W}_{\{j,j\in D_s(i)\}}) & if\ k\ is \\ \prod_{j'\in D_m(i)} s(d_{j'}|k_{t_{\cdot,j'}}^{-i}, k_{t_{i,j'}} = k) & previously\ used \\ \alpha_g f_{k^{new}}^{-\boldsymbol{W}_{j,j\in D_s(i)}}(\boldsymbol{W}_{j,j\in D_s(i)}) & if\ k = k^{new} \\ \prod_{j'\in D_m(i)} s(d_{j'}|k_{t_{\cdot,j'}}^{-i}, k_{t_{i,j'}} = k^{new}) & \end{cases}$$

4.3 Model Analysis

Computation Complexity. The computation burden mainly lies in the Gibbs sampling inference part. Due to the sequential sampling process for the unobserved variables $\boldsymbol{t}$, $\boldsymbol{k}$, and $\boldsymbol{d}$, the computation complexity of Gibbs sampling is linear to the total number of the hidden variables. The size of $\boldsymbol{d}$ is the number of documents in $C_1 = D_m(\cdot)$, The size of $\boldsymbol{t}$ is the number of edges in the bipartite link graph $C_2 = \sum_{j=1}^{N} U(j)$. The size of $\boldsymbol{k}$ can be estimated by its expectation $C_3 = \sum_{i=1}^{M} \alpha_l \log(D_m(i) \cup D_s(i))$. Thus, the computation complexity is about $O(C_1 + C_2 + C_3)$.

Parameter Setting Analysis. Although the model is nonparametric, there are still three parameters, two DP concentration parameters α_l and α_g, and one parameter β in the Dirichlet distribution that is set as the base measure for the global DP. The three parameters can be divided to two groups. α_l and α_g control the prior probability of generating new community. Since we have no prior information about the number of communities, they are set to small values, such as $0.1 \sim 1$ for $alpha_l$, $1 \sim 10$ for $alpha_g$. β controls the impact from the text contents. In typical topic modeling studies, β is always set to around 0.01 so that the text contents can dominate the topic modeling process. While in our setting, β actually controls the balance between the information of text contents and the information of graph structure. The smaller β is the larger the impact of text contents is. We seek to reduce some impact of the text contents so that the graph structure can play a more important role. Therefore, we should set β to a little larger than the usual setting 0.01. From the experiments, we find that $0.01 \sim 0.05$ is a suitable interval.

Parameter Estimation. After model training, there may be two sets of mixture components. One set of mixture components are referred by the documents, i.e. either $d_j, j \in D_m(\cdot)$ or $t_{\cdot,j}, j \in D_s(\cdot)$ will refer to them. We call them MC_{ref} for convenience. Accordingly, the others have no referred documents. We call them MC_{nref}. MC_{ref} is influenced by both document contents and graph structure. While MC_{nref} has no influence on the contents. It is originated from the setting that HDP is placed solely on the graph structure, but the text contents information may only agree with the graph information on a subset of the mixture components brought in by HDP. Since MC_{nref} contains no text information, we drop them. The mixture components in MC_{ref} are remained as the final community space. They can be reconstructed by the posterior expectation of them as

$$\hat{\phi}_k = \int \phi_k p(\phi_k|\{\boldsymbol{W}_j| \begin{cases} d_j = k, j \in D_m(\cdot) \\ t_{\cdot,j} = k, j \in D_s(\cdot) \end{cases}\})d\phi_k \tag{16}$$

And the community distribution θ_i of a user i is estimated from the community assignment variables $d_j, j \in D_m(i)$ and $t_{i,j}, j \in D_s(i)$ as

$$\hat{\theta}_i = \frac{1}{|D_m(i)|} \sum_{j\in D_m(i)} \delta(\phi_{k_{d_j}}) + \frac{1}{|D_s(i)|} \sum_{j\in D_s(i)} \delta(\phi_{k_{t_{i,j}}}) \tag{17}$$

Table 2: Statistics of PAPER and NYT

	# of users	# of docs	# of links	# of links/user
PAPER	9415	5308	25034	2.7
NYT	1677	2461	83367	49.7

5. EXPERIMENTS

To test the effectiveness of the proposed ToP model, we apply it to community discovery on two data collections we crawled from the web, a research proceeding corpus and a New York Times news corpus. We perform parameter sensitivity analysis, followed by community semantic analysis. We also compare it with state-of-the-art community discovery algorithms, NCut and NetPLSA, in the experiments.

5.1 Data Collection

We mainly use the research proceeding corpus to evaluate the detailed performance of our model. The dataset contains the abstracts of all papers from 7 research conferences, i.e. ACL, ICML, SIGGRAPH, SIGIR, SIGKDD, SIGMOD, and WWW, between 2005 and 2009. In total, there are 5308 papers and 9415 individual authors. 11913 unique terms appear at least once in the dataset after we preprocess the data by removing common stop words and stemming. We call this corpus PAPER.

ToP model is not limited to discover communities of users. The New York Times news corpus is collected to verify the model's general applicability. We collect a set of companies[1] and their news articles from New York Times. The whole corpus is quite huge. We build a subset of the complete corpus for experiments. The dataset consists of all the articles that mention about at least 3 companies. And hereafter we refer to it as NYT. Table 2 shows some statistics of PAPER and NYT. All the above datasets are publicly available.

5.2 Evaluation Metric

Given two users' community membership distributions, we use the Categorical Clustering Distance(CCD)[25] to compare the quality of these two distributions. This method relates only to users' community membership distributions. So it can be applied to evaluate different algorithms that generate user community distributions with different community dimensions.

Given a test data set consisting of a set of users $U_T = \{u_i|1 \le i \le M\}$ with their ideal community membership distributions $\{\theta_i|1 \le i \le M\}$ and their computed community membership distributions $\{\hat{\theta}_i|1 \le i \le M\}$. The CCD between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ is given by

$$CCD(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \min_{w_{k,j}} \sum_{k=1}^{K} \sum_{j=1}^{J} w_{k,j} \sum_{i=1}^{M} |\theta_{i,k} - \hat{\theta}_{i,j}| \tag{18}$$

subject to $w_{k,j} \ge 0, \sum_{k=1}^{K} w_{k,j} = \frac{1}{J}, \sum_{j=1}^{J} w_{k,j} = \frac{1}{K}$ for all k, j, where K and J are the number of communities in $\boldsymbol{\theta}$

[1] `http://topics.nytimes.com/topics/news/business/companies/index.html`

and $\hat{\theta}$, respectively. Linear programming is used to compute equation (18).CCD is nonnegative and equals zero if and only if the two distributions are identical. The smaller the value is, the closer those two distributions get to each other.

5.3 Ground Truth and Baseline Setting

According to the definition of community in our setting, we build ideal community membership distributions for the above datasets. For the PAPER corpus, we treat each conference as a community and the proportion of the number of papers one author published in each conference as the ideal probability the author belongs to that community. For NYT, those companies are categorized into 10 "sectors" according to New York Times, such as technology, basic material, finance etc. So we treat the categorization as the ideal community membership for those companies.

Two models are selected as baseline for comparison. One is NetPLSA and the other is NCut. Since the two algorithms both run on user link graph, we need to first build such a graph with weights on the links. The weight of a link connecting u_i and u_j is set to be the number of edges that connects u_i and u_j. As these two models are parametric models, the number of communities are set manually. For PAPER, the number of communities of NetPLSA and NCut are both set to be 7, i.e. the number of conferences in the corpus for the reason that these conferences focus on different subjects. This setting method is the same as what the authors of NetPLSA did in their original work. And to NYT, we set the number of communities to be identical to the number of "sectors", i.e., 10, of those companies. For the parameter γ in NetPLSA, we choose the lead-to-optimal-result one for 9 different settings raging from 0.1 to 0.9 with step 0.1.

5.4 Parameter Sensitivity Analysis

We examine different combinations of α_l, α_g and β in the experiment. For each combination of the parameters, we run the inference for 100 iterations. We present the influence of α_l, α_g and β in Figure 5.

From Figure 5, we see that the number of communities detected by our model is not quite sensitive to parameter changes. The number of communities detected from both PAPER and NYT are around 8. Section 5.6 gives explanations why they are not precisely equal to the ideal community number(mainly because of noise). Though not exactly perfect, the number of communities our model detects show the ability of automatically determining the proper number of communities.

5.5 Community Membership Evaluation

On the PAPER dataset, the evaluation result of NCut $CCD = 1351.02$, and we list the evaluation results of NetPLSA regarding to its parameter λ in Table 3. As we see from Table 3, the best CCD value of NetPLSA is 1408.9, which outperforms NCut, so from now on we choose the best evaluation result of NetPLSA as comparison baseline for ToP. In our experiments, we also test different sets of parameters of ToP with α_l varying from 0.1 to 0.5, α_g from 1.0 to 5.0 and β from 0.01 to 0.20. We show part of the detailed evaluation results of ToP with respect to α_l, α_g and β in Figure 6.

With $\alpha_l = 0.2$, $\alpha_g = 5.0$ and $\beta = 0.05$, we get the minimum value of $CCD = 1219.65$ giving a maximum improvement over NetPLSA as

$$\frac{|1219.65 - 1408.9|}{1408.9} \times 100\% = 13.4\%$$

Table 4: Best CCD of NCut, NetPLSA and ToP

	NCut	NetPLSA	ToP
PAPER	1351.02	1408.9	1219.65
NYT	173.9	180.35	141.82

We pick the configuration of α_l, α_g and β which lead to the lowest CCD. Under that configuration, our model detects 8 communities and we make statistics of papers from each conference as shown in Figure 7. Note here that the community membership of a paper is set to be community assignment in the last iteration of Gibbs sampling. We are unable to get this result by using NetPLSA due to its coarse grained modeling of the documents. The horizontal axis represents the communities and the vertical axis represents the percentage of papers from each conference. We also list the top 10 terms for each community discovered by ToP and NetPLSA in Table 5 and Table 6, respectively.

While on the NYT dataset, we only show the best CCD results of NCut, NetPLSA and ToP as shown in Table 4. ToP outperforms both NetPLSA and NCut on NYT. Besides, from the results on NYT, we can see that NCut even outperforms NetPLSA for the reason that it constructs a densely-connected graph on NYT and that NetPLSA neglects too much important network clues to discover proper community structures.

Because ToP models more detailed participation information than NetPLSA, ToP takes longer time to output the community distribution results than NetPLSA does. On the PAPER corpus, NetPLSA takes about $5 \sim 10$s per iteration while ToP takes about $10 \sim 30$s per iteration on a daily workstation depending on the parameters.

5.6 Community Semantic Analysis

From Figure 7, we see that our model discovered 6 major communities and 2 minor communities. We examine the 6 major communities first. Considering both the percentage of papers from each conference appearing in each community and the top 10 terms of every community listed in Table 5, it is easy to see that c_1 well corresponds to the information retrieval community, c_2 is closely related to computer graphics, c_3 is mainly about data mining, c_4 covers the database community, c_5 mainly concerns about computer linguistics, c_6 is closely related to machine learning. Note that papers from WWW scatter around several communities, which is quite reasonable because the WWW conference covers topics in information retrieval, data mining, data management, etc. As to the 4 minor communities, c_7 and c_8 both contain only 1 paper. After an investigation of the data, we find that the authors of the above 2 papers have no co-authorship with the rest authors in the dataset and that the contents of these papers are very dissimilar from others. It is the above two reasons that prevent our model from determining the same number of communities as the ideal one. Let us take the paper belonging to c_7 as an example. *The Portinari project: IR helps art and culture*, a demo paper authored by João Candido Portinari in SIGIR'05, demonstrates a cultural project of collecting and distributing all works of the Brazilian artist Candido Portinari and shows how IR helps accomplish this

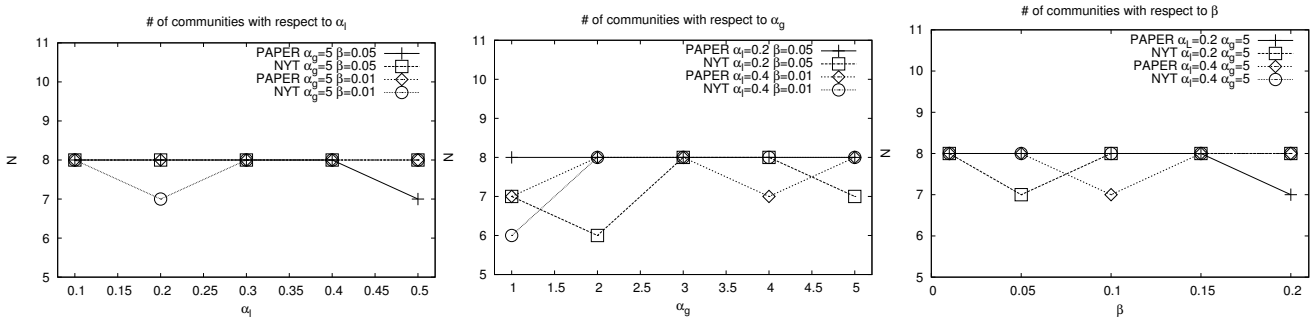

Figure 5: Influence of each parameter on community number

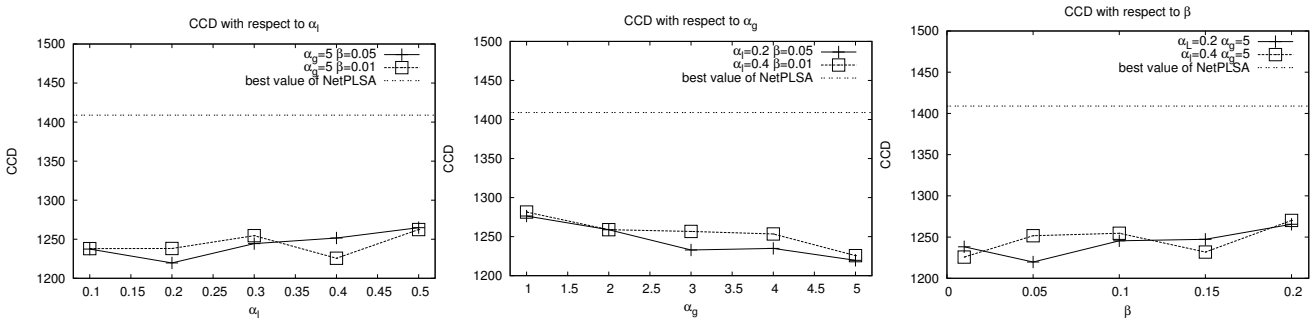

Figure 6: CCD of ToP with respect to each parameter(lower is better)

Table 3: Evaluation result of NetPLSA with respect to the lead-to-optimal-result parameter

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$CCD(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$	1474.2	1463.8	1435.6	1509.9	1486.5	1408.4	1432.6	1435.7	**1408.9**

Table 5: Top 10 terms extracted by ToP for each community

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
search	imag	data	data	model	learn	portinari	economi
web	model	network	queri	base	algorithm	work	engin
queri	method	algorithm	web	languag	model	paint	perspect
user	motion	mine	system	word	data	brazil	understand
inform	base	pattern	servic	method	method	social	long
model	surfac	graph	applic	approach	problem	project	chang
retriev	present	model	databas	translat	cluster	present	arriv
document	mesh	base	base	system	propos	document	econom
base	time	cluster	user	paper	base	develop	largest
result	algorithm	propos	process	show	classif	import	compani

Table 6: Top 10 terms extracted by NetPLSA

c_1	c_2	c_3	c_4	c_5	c_6	c_7
model	data	data	web	search	system	imag
languag	queri	algorithm	network	queri	user	model
word	databas	learn	user	document	interact	method
base	system	model	social	retriev	visual	motion
translat	applic	problem	content	rank	inform	base
method	process	method	page	inform	model	surfac
approach	xml	cluster	commun	result	base	mesh
text	servic	propos	inform	user	present	light
system	perform	set	data	relev	evalu	time
extract	base	approach	mine	base	content	comput

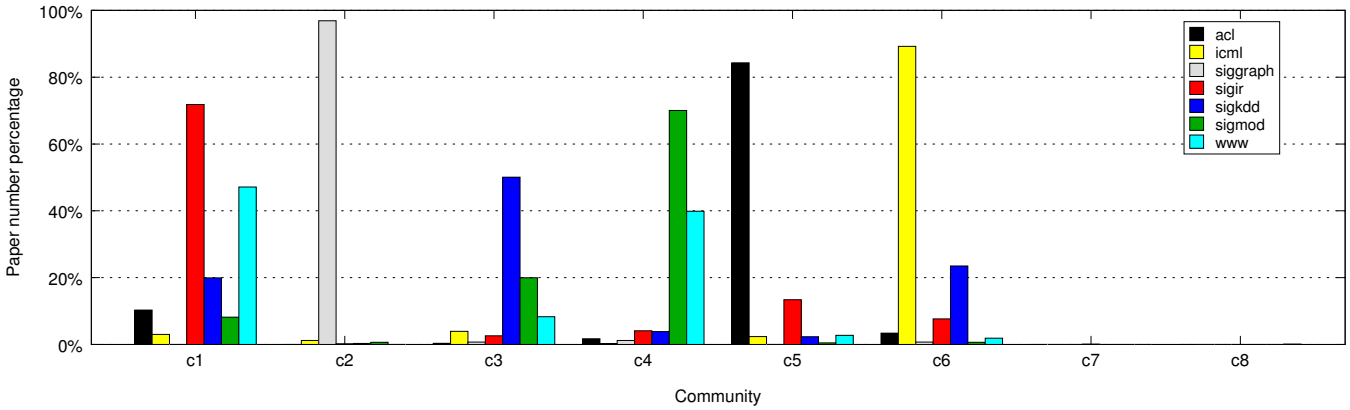

Figure 7: Paper number percentage distribution over community(best viewed in color)

Table 7: Community membership of several active authors

	IR	CG*	DM	DB	CL	ML
W. Bruce Croft	△	△		△	△	
Jamie Callan	△				△	
Chengxiang Zhai	△	△			△	△
James Allan	△	△		△	△	
Andrew McCallum	△				△	△
Kun Zhou		△	△	△		

*CG: Computer Graphics, DB: Database, DM: Data Mining

project and spread the Brazilian culture. The author is actually an isolated point in our co-authorship network. Another example from community c_8, *The New Economy - An Engineer's Perspective*, a keynote by David Brown in WWW'06, tells how mobile communication & computing changes people's daily life and the economy. Also, this author is another isolated point in our co-authorship network.

ToP figures out the community each author gets involved in. Table 7 shows the community membership of several active authors in PAPER. A △ indicates the probability of an author belonging to a community is above some threshold.

From the above analysis, we see that with the number of communities determined automatically, our model has the ability of discovering semantic communities.

5.7 Case Study

This section gives some case studies in the PAPER dataset to show the semantic community discovery ability of ToP. We investigate the PAPER dataset by calculating all the connected components of the co-authorship network and find that there are on average 184 connected components inside the scope of each conference. Also the number of authors who publish papers in more than 2 conferences is 1187. The above two observations imply that the loosely connectd semantic related scenario and densely connected semantic unrelated scenario is common in the PAPER dataset. Thus we show two case studies in each case respectively.

5.7.1 Loosely connected semantic related scenario

For this scenario, we take authors with similar research interests but with neither direct nor indirect co-author relationships to examine our model. That is, they publish paper in the same conference but in different connected components. In this case, we have the following example.

Andrew McCallum and Thomas Hofmann are in two unconnected subgraph in our bipartite graph. Andrew McCallum published nine papers covering topics on data mining in SIGKDD and Thomas Hofmann also contributed one SIGKDD paper *Non-redundant clustering with conditional ensembles* to our dataset. Results of our model successfully assign most of Andrew McCallum's SIGKDD papers and Thomas Hofmann's SIGKDD paper to the same community(c_3 in Figure 7, in which SIGKDD papers make a majority). Therefore, our model can successfully merge loosely connected semantic related communities.

In this scenario, NCut fails to recover this semantic community because it takes no text information into consideration.

5.7.2 Densely connected semantic unrelated scenario

For this seenario, we take authors who have co-author relationships but their co-authored papers are about different research topics to examine our model and we have the following example.

Jerry Scripps, Pang-Ning Tan and Abdol-Hossein Esfahanian together contributed a paper *Measuring the effects of preprocessing decisions and network forces in dynamic network analysis* in SIGKDD'09, while Haibin Cheng and Pang-Ning Tan together contributed a paper *Semi-supervised learning with data calibration for long-term time series forecasting* in SIGKDD'08. It is necessary to mention that there are no records in the PAPER corpus concerning the above 4 authors except these two papers. Therefore, authors of the first paper are densely connected to each other and the same to the authors of the second one. These two author groups share a common vertex (Pang-Ning Tan). Results of ToP assign c_3(Data mining) to the SIGKDD'09 paper and c_6(Machine learning) to the SIGKDD'08 paper, for the reason that the SIGKDD'08 paper mainly focuses on proposing a new learning method. The above analysis shows that our model has successfully split the two densely connected communities whose semantics are not quite related.

While in this scenario, we are unable to recover such detailed community of each paper using NetPLSA due to its coarse grained modeling of the documents by simply merge the above two papers of Pang-Ning Tan into a single document.

6. CONCLUSIONS AND FUTURE WORK

Although community discovery techniques have been developed for decades, there is not much work done in developing general algorithms for u2u-link data while considering textual content information. This paper proposes a principle solution. The main contributions of this paper can be summarized as the following.

1). The proposal of a Topics on Participations(ToP) model for community discovery in u2u-link data. ToP makes it possible for mining detailed information from each participation. Moreover, the proposed model not only captures the structure information induced from the u2u-link graph but also provides a method for automatic selecting proper number of communities.

2). The extensive evaluation on two real-world datasets, which verifies the effectiveness of the proposed model. ToP can automatically detect the proper number of communities, provide a reasonable interpretation of the discovered communities, and outperform the baseline models in mining user community distribution.

There are several potential future directions of this work. First, we will try some other document modeling e.g. PLSA and LDA. Second, we will develop a parallel solution to improve the scalability of our algorithm. Last, by taking temporal dimension into consideration, studying how communities evolve over time also deserves a try.

7. REFERENCES

[1] R. Anderson and K. Lang. Communities from seed sets. In *Proceedings of International Conference on WWW 2006*, pages 475–486, 2006.

[2] C. Andrieun, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2004.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Proc. of NIPS*, 2001.

[5] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. In *Proc. of National Academy of Sciences*, pages 5220–5227, 2004.

[6] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of KDD 2000*, pages 150–160, 2000.

[7] M. Girvan and M. Newman. Community structure in social and biological networks. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 7821–7826, 2002.

[8] C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. Discovering latent classes in relational data, 2004. Technical report.

[9] J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of International Conference on WWW 2010*, pages 631–640, 2010.

[10] H. Li, Z. Nie, W.-C. Lee, C. L. Giles, and J.-R. Wen. Scalable community discovery on textual data with relations. In *Proceedings of CIKM 2008*, pages 1203–1212, 2008.

[11] Y. Lin, Y. Chi, S. Zhe, H. Sundaram, and B. L. Tseng. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proc. of WWW'08*, pages 685–694, 2008.

[12] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: Joint models of topic and author community. In *Proc. of ICML'09*, pages 665–672, 2009.

[13] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.

[14] A. McCallum, X. Wang, and N. Mohanty. Joint group and topic discovery from relations and text. *Statistical Network Analysis: Models, Issues and New Directions*, 2007.

[15] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *Proc. of WWW'08*, pages 101–110, 2008.

[16] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E.*, 69(026113), 2004.

[17] M. E. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proc. Nat. Acad. Sci.*, volume 101, Suppl 1, pages 5200–5205, 2004.

[18] K. Nowicki and T. A. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455), 2001.

[19] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.

[20] J. Ruan and W. Zhang. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In *Proc. of ICDM*, pages 643–648, 2007.

[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transcations of Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[22] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. of SIGKDD'04*, pages 306–315, 2004.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2003.

[24] D. M. Wilkinson and B. A. Huberman. A method for finding communities of related genes. *PNAS*, 101, Suppl 1:5241–5248, 2004.

[25] D. Zhou, J. Li, and H. Zha. A new mallows distance based metric for comparing clusterings. In *Proc. of ICML'05*, pages 1028–1035, 2005.

[26] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. Probabilistic models for discovering e-communities. In *Proc. of WWW'06*, pages 173–182, 2006.