# Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling

ZHIJUN YIN, University of Illinois at Urbana-Champaign
LIANGLIANG CAO, IBM T.J. Waston Research Center
QUANQUAN GU, University of Illinois at Urbana-Champaign
JIAWEI HAN, University of Illinois at Urbana-Champaign

This paper studies the problem of latent community topic analysis in text-associated graphs. With the development of social media, a lot of user-generated content is available with user networks. Along with rich information in networks, user graphs can be extended with text information associated with nodes. Topic modeling is a classic problem in text mining and it is interesting to discover the latent topics in text-associated graphs. Different from traditional topic modeling methods considering links, we incorporate community discovery into topic analysis in text-associated graphs to guarantee the topical coherence in the communities so that users in the same community are closely linked to each other and share common latent topics. We handle topic modeling and community discovery in the same framework. In our model we separate the concepts of community and topic, so one community can correspond to multiple topics and multiple communities can share the same topic. We compare different methods and perform extensive experiments on two real datasets. The results confirm our hypothesis that topics could help understand community structure, while community structure could help model topics.

Categories and Subject Descriptors: H.2.8 [**Database applications**]: Data mining

General Terms: Algorithms

Additional Key Words and Phrases: topic modeling, community discovery

## 1. INTRODUCTION

Topic modeling is a classic text mining task which is to discover the hidden topics that occur in a document collection. Topic models, such as PLSA [Hofmann 1999], LDA [Blei et al. 2003] and their variants [Blei 2011], use a multinomial word distribution to represent a semantic coherent topic and model the generation of the text collection with a mixture of such topics. With more and more text content online, it is difficult for us to read all the documents and digest all the information. Topic modeling provides an effective approach to help understand these huge amounts of information. The discovered topics are also useful to organize and search the content.

With the development of social media, a lot of user-generated content is available with user networks. Users communicate and interact with each other in social media sites. Besides the links among users, users generate a lot of text content as well. Along with rich information in networks, user graphs can be extended with text information on nodes. In social networking sites, users maintain profile pages, write comments and share articles. In photo and video sharing sites, users use short text to tag photos and videos. In micro-blogging sites, users post their status updates. We consider a graph with text on nodes as a text-associated graph.

To discover the community-based latent topics in text-associated graphs, we are interested in the following three tasks. First, we would like to discover the community structure in the graph, so we can know the relationships among different users. Identified communities not only can provide summarization of network structure and help understand the graphs, but also are important to analyze user behaviors in the setting of social networks. Second, we would like to discover the latent topics in text-associated graphs. In this way we can know the interests of the users in the graph. Third, we would like to learn the relationship between communities and topics, so we can know which communities are interested in a specific topic or which topics a specific community cares about.

In this paper we incorporate community discovery into topic analysis and propose a community-based topic analysis framework called LCTA (Latent Community Topic Analysis). With the development of social networks, discovering communities in graphs draws much more attention than before [Parthasarathy et al. 2011]. A community in a network is considered as a group of nodes with more interactions and common topics among its members than between its members and others, and community discovery is the process to group the nodes into the clusters of close interaction and common interests. To discover communities in graphs, typically an objective function is chosen to capture the intuition of a community as a set of nodes with better internal connectivity than external connectivity based on link structure [Leskovec et al. 2010]. However, if we only use link to discover communities, we cannot capture the coherence of common interests inside communities. A good community should be coherent in interaction patterns as well as shared topics. Most of previous studies overlook the connection between interactions and interests, and hence might have difficulties in finding the most appropriate communities. To discover the community-based topics in text-associated graphs, we follow the previous text mining studies [Hofmann 1999; Blei et al. 2003; McCallum et al. 2005; Liu et al. 2009] by using topics to model text corpus. Our work is different from the previous work in our assumption that topic and community are different concepts. Instead of modeling topics by considering pair-wise link relationships, we consider topic modeling in the community level. We assume that one community can correspond to multiple topics and multiple communities can share the same topic. For example, in a network one community can be interested in both politics and entertainment topics, while multiple communities can be interested in a politics topic. The analysis of topics and communities could benefit each other. In our model, users are likely to form a link to another user from the same community and users in the same community usually share coherent interests as topics. Topics are generated from communities in our method, so it captures the topical coherence in the community level. As we will see in the experimental results, the interaction of communities and topics provides flexibility in both community discovery and topic modeling process.

The contributions of the paper are summarized as follows.

— We introduced the problem of latent community topic analysis.
— We proposed a model called LCTA to incorporate community discovery into topic modeling.
— We performed extensive experiments on two real datasets to demonstrate the effectiveness of our LCTA method.

The rest of the paper is organized as follows. We introduce the problem of latent community topic analysis in Section 2. We propose our method LCTA in Section 3. We compare different methods and show the performance in Section 4. We summarize the related work in Section 5 and conclude the paper in Section 6.

## 2. PROBLEM FORMULATION

In this section, we introduce the problem of latent community topic analysis and define the related concepts. The notations used in this paper are listed in Table I.

*Definition* 2.1. A **text-associated graph** is a graph with text information on nodes. Formally, $G(U, E)$ is a graph that contains users and edges, where $U$ is the user set in $G$ and $E$ is the edge set in $G$. $u$ is a user in $U$ that consists of both text and links. $\mathbf{w}_u$ is the text part of user $u$ and $\mathbf{l}_u$ is the link part of user $u$.

Table I. Notations used in the paper

| Symbol | Description |
|---|---|
| $G$ | A graph that contains users and links |
| $U$ | The user set in $G$ |
| $E$ | The edge set in $G$ |
| $V$ | Vocabulary set, $w$ is a word in $V$ |
| $C$ | The community set, $c$ is a community in $C$ |
| $Z$ | The topic set, $z$ is a topic in $Z$ |
| $u$ | A user $u$ that is associated with texts and links |
| $\mathbf{w}_u$ | The text of user $u$ |
| $\mathbf{l}_u$ | The links of user $u$ |
| $\alpha$ | The community distribution set for $U$, i.e., $\{\alpha_u\}_{u\in U}$ |
| $\alpha_u$ | The community distribution for user $u$, i.e., $\{p(c\|u)\}_{c\in C}$ |
| $\theta$ | The word distribution set for $Z$, i.e., $\{\theta_z\}_{z\in Z}$ |
| $\theta_z$ | The word distribution for topic $z$, i.e., $\{p(w\|z)\}_{w\in V}$ |
| $\phi$ | The topic distribution set for $C$, i.e., $\{\phi_c\}_{c\in C}$ |
| $\phi_c$ | The topic distribution for community $c$, i.e., $\{p(z\|c)\}_{z\in Z}$ |
| $\eta$ | The user distribution set for $C$, i.e., $\{\eta_c\}_{c\in C}$ |
| $\eta_c$ | The user distribution for community $c$, i.e., $\{p(v\|c)\}_{v\in U}$ |

*Definition* 2.2. A **community** is a group of users in the graph with more interactions and common interests within the group than between groups. We denote $C$ as the community set and $c$ is a community in $C$.

The conditional probability of a community given a user represents the participation level of the user in the community. Formally, $p(c|u)$ is the probability of community $c$ given user $u$, s.t., $\sum_{c\in C} p(c|u) = 1$. We denote $\alpha$ as the community distribution set for user set $U$, i.e.,$\{\alpha_u\}_{u\in U}$ where $\alpha_u = \{p(c|u)\}_{c\in C}$. From $\alpha_u$, we can infer to which community user $u$ is most likely to belong.

*Definition* 2.3. A **topic** is a semantically coherent theme, which is represented by a multinomial distribution of words. Formally, each topic $z$ is represented by a word distribution $\theta_z = \{p(w|z)\}_{w\in V}$ s.t. $\sum_{w\in V} p(w|z) = 1$. We denote $Z$ as the topic set and $\theta$ as the word distribution set for $Z$, i.e., $\{\theta_z\}_{z\in Z}$.

The conditional probability of a topic given a community represents the relationship between the topic and the community. Formally, $p(z|c)$ is the probability of topic $z$ given user $c$, s.t., $\sum_{z\in Z} p(z|c) = 1$. We denote $\phi$ as the topic distribution set for $C$, i.e., $\{\phi_c\}_{c\in C}$ where $\phi_c = \{p(z|c)\}_{z\in Z}$. From $\phi_c$, we can infer which topic community $c$ is mostly interested in.

To help understand the above definitions, we give two examples below.

*Example* 2.4. In *DBLP*[1] (a digital computer science bibliographic graph), authors are considered as users, the paper titles of the authors are the text of users and the co-authorship relationship forms the links between users. In this text-associated graph, communities are the groups of authors that have close collaboration and common research interests with each other, and topics can be different research areas in computer science domains.

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

*Example* 2.5. In *Twitter*[2] (a micro-blogging site), users can post text of up to 140 characters on their profile pages. The published tweets are the text of users and the follower relationship forms the links of users. In this text-associated graph, communities are the groups of users that have similar follower patterns and common discussed topics with each other, and topics can be the popular themes in the social community.

Given the definitions of text-associated graph, community and topic, we define the problem of latent community topic analysis as follows.

*Definition* 2.6. **Latent community topic analysis** is the process to group the nodes in a graph into different communities and discover the topics that are coherent in communities. Formally, given a text-associated graph $G(U, E)$, the number of communities $N$ and the number of topics $K$, we would like to know the following results in latent community topic analysis.

— The community distribution set $\alpha$ for user set $U$, i.e., $\{\alpha_u\}_{u\in U}$ where $\alpha_u$ is the community distribution for user $u$, i.e., $\{p(c|u)\}_{c\in C}$ where $|C|$=$N$. Based on $\alpha$ we can assign users to the most likely communities that they belong to.

— The word distribution set $\theta$ for topic set $Z$, i.e., $\{\theta_z\}_{z\in Z}$ where $|Z|$=$K$ and $\theta_z$ is the word distribution for topic $z$, i.e., $\{p(w|z)\}_{w\in V}$. Based on $\theta$ we can know the the discussed topics in the text-associated graph.

— The topic distribution set $\phi$ for community set $C$, i.e., $\{\phi_c\}_{c\in C}$ where $\phi_c$ is the topic distribution for community $c$, i.e., $\{p(z|c)\}_{z\in Z}$. Based on $\phi$, we can know the relationship between topics and communities, i.e., which topics are related to a specific community.

## 3. LATENT COMMUNITY TOPIC ANALYSIS

In this section we introduce our framework of latent community topic analysis. First, we propose a model called LCTA. Second, we introduce how to estimate the parameters in the model. Third, we analyze the complexity of our algorithm.

### 3.1. General Idea

In our LCTA (Latent Community Topic Analysis) model, we would like to discover both communities and topics in a text-associated graph. The network structure provides information of how popular a node is and how it is connected to its neighbor nodes. When a group of nodes are closely connected together and share common interests, the group can be considered as a community. Another important information existing in a text-associated graph is topic. If we explore the semantics of the text in the graph, we can find meaningful topics shared by different users.

In LCTA we use the following characteristics in text-associated graphs.

— Topic and community are different concepts. In practice, a community can incorporate multiple topics, while multiple communities can share the same topic. For example, in DBLP (a digital computer science bibliographic graph), one community can be interested in both “database” and “data mining” topics, while multiple communities can be interested in “information retrieval” topic.

— Good community structure is useful for modeling topics. The users in the same community are closely connected to each other and share common topics. Topics are related to community structure instead of individual nodes. Therefore, we can guarantee the topic coherence in the community level. For example, in DBLP, if several

[2]http://twitter.com/

authors often collaborate together and form a community, we can assume that they have some common research interests. The discovered community structure can be used for generating more meaningful topics.

— Meaningful topics can help discover communities. The users form a community not only because they are linked to each other but also because they share common topics. Instead of single terms, topics are used as the latent concepts in the text, which can represent different aspects more comprehensively. Meaningful topics can guide the discovery of community structure. For example, if two authors are both interested in "data mining" topic, they are more likely to belong to the same community. Besides, the link graph may not be complete sometimes, so the topics can provide additional information for community discovery.

Based on the above characteristics in text-associated graphs, we would like to integrate community discovery and topic modeling in our LCTA model. It is not difficult to see that the analysis of communities and topics can mutually enhance each other. A good community needs to be coherent in both links and topics. A user is more likely to form a link with another user within the same community and the users in the same community share the common topics. We believe that integrating both community structure and text topics will lead to a better description of communities and hence more accurate analysis. To model topics, we consider the topical coherence in the community level beyond the constraints on the pair-wise relationship. In traditional topic modeling methods topics are from documents and the relationship between terms and documents are predetermined. In our model topics are from communities and we explore the relationship between terms and communities, so communities in our method can be considered as pseudo-documents. The communities in our model are not predefined and they are discovered along with the topic modeling process. Through the mutual enhancement between community discovery and topic modeling, we can discover the communities that are coherent in both link and topical structure and identify the topics that are coherent in the community level.

### 3.2. Generative Process in LCTA

The generative process to generate a text-associated graph is as follows.

For each user $u$ in a text-associated graph $G$:

(1) To generate each word for user $u$:
   (a) Sample a community $c$ from multinomial $\alpha_u$.
   (b) Sample a topic $z$ from multinomial $\phi_c$.
   (c) Sample a word $w$ from multinomial $\theta_z$.
(2) To generate each link for user $u$:
   (a) Sample a community $c$ from multinomial $\alpha_u$.
   (b) Sample a user $v$ from multinomial $\eta_c$ and form a link between user $u$ and $v$.

In order to generate a user $u$ in graph $G$, we need to generate both the text $\mathrm{w}_u$ and the links $\mathrm{l}_u$ of user $u$. To generate each word in $\mathrm{w}_u$, we first sample a community $c$ from multinomial $\alpha_u$. $\phi_c$ is the topic distribution for community $c$. Since the users in the same community are likely to have the same topics, we sample a topic $z$ from $\phi_c$. Lastly we sample a word $w$ from multinomial $\theta_z$. To generate each link in $\mathrm{l}_u$, we first sample a community $c$ from multinomial $\alpha_u$. $\eta$ is the user distribution set for community set $C$, i.e., $\{\eta_c\}_{c\in C}$ where $\eta_c$ is $\{p(v|c)\}_{v\in U}$. $\eta_c$ can be considered as the user participation in community $c$. Since a user is more likely to link to another user from

the same community, we sample a user $v$ from multinomial $\eta_c$ and form a link between user $u$ and $v$.

Given the data collection $\{(\mathbf{w}_u, \mathbf{l}_u)\}_{u \in U}$ where $\mathbf{w}_u$ represents the text of user $u$ and $\mathbf{l}_u$ represents the links of user $u$, the log-likelihood of the collection is as follows.

$$\begin{aligned} L(G) &= \log \prod_{u \in U} p(\mathbf{w}_u, \mathbf{l}_u) \\ &= \sum_{u \in U} \log p(\mathbf{w}_u, \mathbf{l}_u) \\ &\propto \sum_{u \in U} \log \sum_{c \in C} p(\mathbf{w}_u|c) p(c|u) \sum_{c \in C} p(\mathbf{l}_u|c) p(c|u) \end{aligned}$$

$$p(\mathbf{l}_u|c) = \prod_{(u,v) \in E} p(v|c) \tag{1}$$

$$p(\mathbf{w}_u|c) = \prod_{w \in \mathbf{w}_u} p(w|c) \tag{2}$$

We assume that the words in each community are generated from a mixture of a background model and the community-based topic models. The purpose of using a background model is to make the topics concentrated more on discriminative words, which leads to more informative models [Zhai et al. 2004].

$$p(w|c) = \lambda_B p(w|B) + (1 - \lambda_B) \sum_{z \in Z} p(w|z) p(z|c) \tag{3}$$

$\lambda_B$ is the mixing weight of the background model and a large $\lambda_B$ can exclude the common words from the topics. In this paper the mixing weight $\lambda_B$ is set as 0.9 following the empirical studies [Zhai et al. 2004; Mei et al. 2006]. $p(w|B)$ is the background model, which we set as follows.

$$p(w|B) = \frac{\sum_{u \in U} n(u, w)}{\sum_{w \in V} \sum_{u \in U} n(u, w)} \tag{4}$$

where $n(u, w)$ is the frequency of word $w$ with regard to user $u$.

### 3.3. Parameter Estimation

In order to estimate parameters $\alpha, \theta, \phi, \eta$ in log-likelihood, we use maximum likelihood estimation. In particular, we use Expectation Maximization (EM) algorithm to estimate parameters, which iteratively computes a local maximum of likelihood. In the EM algorithm, we introduce the probabilities of the hidden variables, i.e., $p(c, z|u, w)$ and $p(c|u, v)$ where $p(c, z|u, w)$ is the probability of word $w$ in user $u$ belonging to community $c$ and topic $z$ and $p(c|u, v)$ is the probability of linked user $v$ in terms of user $u$ belonging to community $c$. In the E-step, it computes the expectation of the complete likelihood. In the M-step, it finds the estimation of the parameters that maximizes the expectation of the complete likelihood.

In the **E-step**, $p(c, z|u, w)$ and $p(c|u, v)$ are updated according to Bayes formulas.

$$p(c, z|u, w) \leftarrow \frac{(1 - \lambda_B) p(w|z) p(z|c) p(c|u)}{\lambda_B p(w|B) + (1 - \lambda_B) \sum_{c' \in C} \sum_{z' \in Z} p(w|z') p(z'|c') p(c'|u)} \tag{5}$$

$$p(c|u,v) \leftarrow \frac{p(v|c)p(c|u)}{\sum_{c' \in C} p(v|c')p(c'|u)} \tag{6}$$

In the **M-step**, we update the parameters as follows, where $n(u,w)$ is the frequency of word $w$ with regard to user $u$ and $n(u,v)$ is the weight of the link from user $u$ to $v$.

$$p(z|c) \leftarrow \frac{\sum_{u \in U} \sum_{w \in V} n(u,w)p(c,z|u,w)}{\sum_{u \in U} \sum_{w \in V} \sum_{z' \in Z} n(u,w)p(c,z'|u,w)} \tag{7}$$

$$p(w|z) \leftarrow \frac{\sum_{u \in U} \sum_{c \in C} n(u,w)p(c,z|u,w)}{\sum_{u \in U} \sum_{c \in C} \sum_{w' \in V} n(u,w')p(c,z|u,w')} \tag{8}$$

$$p(v|c) \leftarrow \frac{\sum_{u \in U} n(u,v)p(c|u,v)}{\sum_{u \in U} \sum_{(u,v') \in E} n(u,v')p(c|u,v')} \tag{9}$$

$$p(c|u) \leftarrow \frac{\sum_{z \in Z} \sum_{w \in V} n(u,w)p(c,z|u,w) + \sum_{v \in U} n(u,v)p(c|u,v)}{\sum_{c' \in C} \sum_{z \in Z} \sum_{w \in V} n(u,w)p(c',z|u,w) + \sum_{c' \in C} \sum_{v \in U} n(u,v)p(c'|u,v)} \tag{10}$$

We update topic-related parameters $p(z|c)$ and $p(w|z)$ in Equations 7 and 8 and update community-related parameter $p(v|c)$ in Equation 9. In Equation 10, the community distribution of a user $p(c|u)$ is updated according to the information from both topics and links. The EM steps can be considered as the mutual enhancement between community discovery and topic modeling. In our model, topics are generated from communities, so a good community grouping can help extract meaningful topics. On the other hand, since the communities are coherent in topics, a good topic modeling can improve community discovery process.

### 3.4. Complexity Analysis

We analyze the complexity of parameter estimation process in this section. In the E-step, it needs $O(KN|W|)$ to calculate $p(c,z|u,w)$ in Equation 5 for all $(u,w)$ pairs, where $K$ is the number of topics, $N$ is the number of communities and $|W|$ is the number of words in all the users. To calculate $p(c|u,v)$ in Equation 6 for all $(u,v)$ pairs, it needs $O(N|E|)$ where $|E|$ is the number of edges in the graph. In the M-step, it needs $O(KN|W|)$ to update $p(z|c)$ in Equation 7 for all the communities and to update $p(w|z)$ in Equation 8 for all the topics. It needs $O(N|E|)$ to update $p(v|c)$ in Equation 9 for all the communities. To get updated $p(v|c)$ in Equation 9, it needs $O(KN|W| + N|E|)$. Therefore, the time complexity is $O(iter(KN|W| + N|E|))$ where $iter$ is the number of iterations in the EM algorithm. To store $p(c,z|u,w)$, $p(c|u,v)$, $p(z|c)$, $p(w|z)$, $p(v|c)$ and $p(c|u)$ for all the possible pairs, it needs $O(KN|W|)$, $O(N|E|)$, $O(KN)$, $O(K|V|)$, $O(N|U|)$ and $O(N|U|)$ respectively, where $|V|$ is the size of vocabulary and $|U|$ is the number of users. In total, the space complexity is $O(KN|W| + N|E|)$. EM algorithm can be parallelized with MapReduce [Lin and Dyer 2010], so our algorithm is scalable to large-scaled datasets.

## 4. EXPERIMENT

In this section, we demonstrate the evaluation results of our method. First, we introduce the datasets used in the experiment. Second, we demonstrate the discovered topics and the corresponding communities by our LCTA (Latent Community Topic Analysis) model. Third, we compare our method with other community discovery methods.

Fourth, we compare our method with other topic modeling methods. Lastly we study the effect of parameter changes on the results.

### 4.1. Datasets

We evaluate the proposed method on two datasets – DBLP and Twitter.

— *DBLP* Digital Bibliography Project (DBLP) is a computer science bibliography. We collected the authors in four categories including data mining, databases, machine learning and information retrieval according to the labeling in [Gao et al. 2009]. In this data set, authors are considered as users, the paper titles of the authors are the text of users and the co-authorship relationship forms the links of users. There are 4236 users, 5577 unique terms and 15272 links.

— *Twitter* Twitter is a micro-blogging site where users can post text of up to 140 characters on their profile pages. We collected the tweets related to "obama" and "social media" published by the users from the celebrity list in [Kwak et al. 2010]. In this data set, the tweets are the text of users and the follower relationship forms the links of users. There are 1023 users, 5361 unique terms and 350929 links.

### 4.2. Topics and Communities Discovered by LCTA

In this section, we demonstrate the discovered topics and the corresponding communities by our LCTA (Latent Community Topic Analysis) model in the datasets.

*4.2.1. DBLP.* In DBLP dataset, we set the number of communities as 20 and the number of topics as 4. The topics are listed in Table II. From the result in Table II, we can see that our method can discover the topics in four different areas. Topic 1 is about information retrieval and its popular words are *information*, *retrieval*, *web*, *search*, *query*, *document*, etc. Topic 2 focuses on the words like *learning*, *classification* and *reasoning*. We can infer that it is about machine learning. Topic 3 is about data mining with its emphasis on *mining*, *clustering*, *frequent*, *patterns*, etc. Topic 4 is related to database and its popular words contain *database*, *query*, *system* and *xml*.

Table II. Topics by LCTA in DBLP dataset

| Topic 1 (information retrieval) | Topic 2 (machine learning) | Topic 3 (data mining) | Topic 4 (database) |
|---|---|---|---|
| retrieval 0.0425 | learning 0.0393 | data 0.0451 | data 0.0350 |
| information 0.0329 | knowledge 0.0095 | mining 0.0396 | database 0.0274 |
| web 0.0328 | classification 0.0086 | efficient 0.0174 | query 0.0184 |
| search 0.0234 | reasoning 0.0084 | clustering 0.0159 | system 0.0157 |
| text 0.0205 | model 0.0079 | databases 0.0130 | xml 0.0149 |
| query 0.0155 | analysis 0.0068 | time 0.0102 | databases 0.0144 |
| document 0.0155 | models 0.0067 | large 0.0102 | systems 0.0137 |
| language 0.0115 | approach 0.0064 | patterns 0.0097 | queries 0.0129 |
| user 0.0094 | algorithm 0.0062 | frequent 0.0087 | management 0.0117 |
| system 0.0090 | planning 0.0061 | queries 0.0085 | object 0.0104 |

In Table IV, we show the selected discovered communities related to four different topics and their user distributions inside the communities. As we can see from the result in Table IV, one topic can correspond to multiple communities and the authors in the same community are closely related to each other. Besides, in our model one community can be related to multiple topics. For example, in our experiment one community has the probability 70.21% in data mining topic and 29.79% in database topic, and its users include Wei Wang 0.1159, Jeffrey Xu Yu 0.0966, Hongjun Lu 0.0929, Haixun

Table III. Topics by LCTA in Twitter dataset

| Topic 1 (Obama) | | Topic 2 (social media) | |
|---|---|---|---|
| president 0.0064 | michelle 0.0032 | social 0.0153 | video 0.0043 |
| health 0.0056 | administration 0.0031 | media 0.0144 | sites 0.0039 |
| care 0.0052 | bill 0.0031 | marketing 0.0098 | tools 0.0039 |
| obama 0.0051 | afghanistan 0.0031 | twitter 0.0089 | web 0.0038 |
| barack 0.0046 | bush 0.0029 | business 0.0077 | online 0.0038 |
| speech 0.0044 | nobel 0.0029 | blog 0.0056 | brand 0.0037 |
| reform 0.0038 | peace 0.0027 | facebook 0.0049 | company 0.0035 |
| white 0.0038 | calls 0.0026 | ways 0.0046 | roi 0.0032 |
| house 0.0037 | poll 0.0026 | post 0.0046 | expert 0.0030 |
| plan 0.0035 | iran 0.0026 | tips 0.0045 | guide 0.0030 |

Wang 0.0676, etc. It is a representative community whose members are interested in both data mining and database.

*4.2.2. Twitter.* In Twitter dataset, we set the number of communities as 20 and the number of topics as 2. In Table III, we listed two topics. Topic 1 is about Obama with its focus on *health*, *care*, *white*, *house*, etc. Topic 2 is about social media. In Twitter, many popular users are entrepreneurs and marketers, so its popular words in Topic 2 contain *social*, *media*, *marketing* and *business*. Some users are technology lovers and some specialize in development and search engine optimization, so the words like *ways* and *tips* are also popular.

We show several selected communities and their user distributions in Table V. In the communities related to topic Obama, Community 1 is about news media. For example, NPR Politics is political coverage and conversation from NPR News. NewsHour is one of the most trusted news programs on TV. David Shuster is a journalist for NBC News and MSNBC. Karl Rove is the former deputy chief of staff to President George W. Bush and is the author of Courage and Consequence. Community 2 is a community about conservative politics. For example, MichaelPatrick Leahy is the author of Rules for Conservative Radicals. ChadTEverson is a conservative activist. Nansen Malin is a student who does conservative politics. Most of the users in the communities related to social media are entrepreneurs, strategists, authors, speakers, business coaches, etc. For example, Zee M Kane is the editor-in-chief of The Next Web. Robert Clay is an entrepreneur and business mentor to aspiring market leaders. Jonathan Nafarrete is a social media strategist.

### 4.3. Comparison with Community Discovery Methods

Our LCTA (Latent Community Topic Analysis) model is closely related to community discovery. LCTA can handle community discovery and topic modeling simultaneously. Specifically, in LCTA topics are generated from different communities. In this way we guarantee topical coherence in the community level. It is interesting to compare the performance of our model with community discovery methods.

We compare the following methods in this section.

— NormCut [Shi and Malik 2000]: Normalized cut algorithm on a link graph
— SSNLDA [Zhang et al. 2007]: An LDA-based hierarchical Bayesian algorithm on a link graph where communities are modeled as latent variables and defined as distributions over user space

Table IV. Selected communities by LCTA in DBLP dataset

| Selected communities related to Topic 1 (information retrieval) | |
|---|---|
| Community 1 | Community 2 |
| Clement T. Yu 0.0609 | Charles L. A. Clarke 0.0450 |
| W. Bruce Croft 0.0586 | Susan T. Dumais 0.0420 |
| Abdur Chowdhury 0.0281 | Stefan Buttcher 0.0270 |
| Mark Sanderson 0.0275 | ChengXiang Zhai 0.0217 |
| Aya Soffer 0.0258 | Ryen W. White 0.0195 |
| David Carmel 0.0246 | Jaime Teevan 0.0165 |
| Michael Herscovici 0.0176 | Gareth J. F. Jones 0.0135 |
| Yoelle S. Maarek 0.0176 | Takenobu Tokunaga 0.0135 |
| Steven M. Beitzel 0.0164 | Alvaro Barreiro 0.0120 |
| Andrei Z. Broder 0.0140 | Scott B. Huffman 0.0120 |
| Selected communities related to Topic 2 (machine learning) | |
| Community 1 | Community 2 |
| Andrew McCallum 0.0491 | Tao Li 0.0598 |
| William C. Regli 0.0278 | Changshui Zhang 0.0484 |
| Raymond J. Mooney 0.0278 | Fei Wang 0.0385 |
| Adnan Darwiche 0.0256 | Michael H. Bowling 0.0228 |
| Evan Sultanik 0.0235 | Andrew W. Moore 0.0208 |
| Kiri Wagstaff 0.0214 | Shenghuo Zhu 0.0199 |
| Naoki Abe 0.0199 | Lawrence Birnbaum 0.0171 |
| Rayid Ghani 0.0192 | Jeffrey Junfeng Pan 0.0171 |
| Ian Davidson 0.0186 | Peter Stone 0.0171 |
| Chidanand Apte 0.0176 | James T. Kwok 0.0142 |
| Selected communities related to Topic 3 (data mining) | |
| Community 1 | Community 2 |
| Jiawei Han 0.1594 | Christos Faloutsos 0.1625 |
| Ke Wang 0.0551 | Spiros Papadimitriou 0.0542 |
| Xifeng Yan 0.0541 | Jimeng Sun 0.0362 |
| Hong Cheng 0.0315 | H. V. Jagadish 0.0314 |
| Osmar R. Zaiane 0.0218 | Yaron Kanza 0.0186 |
| Martin Ester 0.0197 | Anand Rajaraman 0.0180 |
| Wen Jin 0.0162 | Hiroyuki Kitagawa 0.0170 |
| Mohammed Javeed Zaki 0.0138 | Caetano Traina Jr. 0.0155 |
| Nick Cercone 0.0118 | Flip Korn 0.0113 |
| Feida Zhu 0.0118 | Lise Getoor 0.0105 |
| Selected communities related to Topic 4 (database) | |
| Community 1 | Community 2 |
| Michael Stonebraker 0.0713 | Jennifer Widom 0.1089 |
| Dirk Van Gucht 0.0282 | Renee J. Miller 0.0528 |
| Jan Van den Bussche 0.0282 | Lucian Popa 0.0409 |
| Samuel Madden 0.0265 | Ronald Fagin 0.0326 |
| Martin Theobald 0.0224 | Laura M. Haas 0.0287 |
| Jan Paredaens 0.0216 | David B. Lomet 0.0255 |
| Patrick E. O'Neil 0.0216 | Zachary G. Ives 0.0215 |
| Elizabeth J. O'Neil 0.0182 | Val Tannen 0.0204 |
| Jose A. Blakeley 0.0166 | Yannis Velegrakis 0.0187 |
| Lipyeow Lim 0.0163 | Chris Clifton 0.0166 |

Table V. Selected communities by LCTA in Twitter dataset

| Selected communities related to Topic 1 (Obama) | |
|---|---|
| Community 1 | Community 2 |
| NPR Politics 0.0249 | Tabitha Hale 0.0093 |
| Paula Poundstone 0.0207 | MichaelPatrick Leahy 0.0075 |
| Los Angeles Times 0.0166 | Infidels Are Cool 0.0068 |
| NPR News 0.0142 | It's Only Words 0.0064 |
| Redeye Chicago 0.0135 | Justin Hart 0.0061 |
| Jim Long 0.0134 | michaelemlong 0.0061 |
| NewsHour 0.0107 | ChadTEverson 0.0058 |
| Steve Garfield 0.0103 | Nansen Malin 0.0057 |
| David Shuster 0.0101 | Markham Robinson 0.0055 |
| Karl Rove 0.0092 | Andrew Windham 0.0055 |
| Selected communities related to Topic 2 (social media) | |
| Community 1 | Community 2 |
| Social Media Insider 0.0062 | Robert Clay 0.0092 |
| Jeff Flowers 0.0061 | Jonathan Nafarrete 0.0091 |
| Wolfgang Jaegel 0.0059 | Andrew Windham 0.0091 |
| Zee M Kane 0.0059 | Arleen Boyd 0.0083 |
| Mark Fulton 0.0056 | Bobby Bloggeries 0.0077 |
| Social Media News 0.0054 | Montaignejns 0.0076 |
| Nick Donnelly 0.0052 | Glen Gilmore 0.0074 |
| Brian Tercero 0.0048 | Larry Brauner 0.0071 |
| Alex Blom 0.0045 | Guy Kawasaki 0.0071 |
| Monik Pamecha 0.0044 | Jay Oatway 0.0070 |

— LCTA: Our Latent Community Topic Analysis model that integrates community discovery with topic modeling

To compare the discovered topics from LCTA with the ones based on NormCut and SSNLDA, we first use NormCut and SSNLDA to cluster the link graph into 20 communities in both datasets and pool the text of the users in the same community together. We consider the text in each community as a document and run topic modeling method PLSA [Hofmann 1999] on the collection. We set the number of topics as 4. NormCut+PLSA and SSNLDA+PLSA can be considered as the approaches to discover topics based on clustered communities. The topics discovered by NormCut+PLSA and SSNLDA+PLSA in DBLP dataset are listed in Table VI and Table VII. We can see the topics discovered by NormCut+PLSA and SSNLDA+PLSA are not meaningful and the result by LCTA in Table II is much better. The topics discovered by NormCut+PLSA and SSNLDA+PLSA in Twitter dataset are listed in Table VIII. We can see that compared with the result by LCTA in Table III, the topics discovered by these two approaches are not pure enough. In Table VIII, the topic related to Obama contains the terms like *social media*, while the topic related to social media also contains the term *obama*. Therefore we can see that our LCTA model considering community discovery and topic modeling in a unified framework performs better than those approaches processing community discovery and topic modeling separately.

Besides the qualitative evaluation, we also quantitatively evaluate the topical coherence. In DBLP dataset, each user is categorized into one domain of data mining, databases, machine learning and information retrieval according to the labeling in [Gao et al. 2009]. Therefore, accuracy and normalized mutual information (NMI) [Cai et al. 2008] can be used to measure the clustering performance and topical coherence in DBLP dataset.

Table VI. Topics by NormCut+PLSA in DBLP dataset

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| data 0.0241 | data 0.0326 | learning 0.0215 | retrieval 0.0237 |
| mining 0.0171 | database 0.0142 | web 0.0163 | information 0.0167 |
| clustering 0.0109 | query 0.0139 | search 0.0108 | data 0.0141 |
| system 0.0095 | queries 0.0118 | mining 0.0102 | learning 0.0114 |
| learning 0.0091 | databases 0.0107 | data 0.0100 | search 0.0108 |
| databases 0.0089 | xml 0.0105 | query 0.0069 | system 0.0106 |
| efficient 0.0087 | efficient 0.0097 | information 0.0068 | web 0.0098 |
| information 0.0086 | mining 0.0086 | classification 0.0066 | query 0.0088 |
| database 0.0085 | web 0.0085 | model 0.0065 | text 0.0085 |
| approach 0.0076 | system 0.0082 | approach 0.0063 | document 0.0073 |

Table VII. Topics by SSNLDA+PLSA in DBLP dataset

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| data 0.0325 | data 0.0248 | data 0.0286 | data 0.0243 |
| database 0.0148 | learning 0.0138 | learning 0.0123 | mining 0.0199 |
| query 0.0132 | retrieval 0.0124 | mining 0.0114 | web 0.0154 |
| system 0.0111 | information 0.0111 | query 0.0109 | search 0.0130 |
| learning 0.0099 | query 0.0093 | information 0.0102 | learning 0.0123 |
| databases 0.0098 | mining 0.0093 | system 0.0100 | information 0.0112 |
| queries 0.0096 | database 0.0093 | database 0.0096 | retrieval 0.0112 |
| systems 0.0094 | system 0.0092 | databases 0.0093 | query 0.0103 |
| efficient 0.0089 | search 0.0079 | web 0.0085 | databases 0.0102 |
| clustering 0.0088 | efficient 0.0078 | retrieval 0.0082 | efficient 0.0101 |

Table VIII. Topics by NormCut+PLSA and SSNLDA+PLSA in Twitter dataset

| NormCut+PLSA | | SSNLDA+PLSA | |
|---|---|---|---|
| Topic 1 | Topic 2 | Topic 1 | Topic 2 |
| obama 0.0103 | media 0.0099 | obama 0.0102 | media 0.0145 |
| president 0.0065 | social 0.0096 | president 0.0067 | social 0.0141 |
| health 0.0047 | marketing 0.0061 | health 0.0054 | obama 0.0103 |
| care 0.0047 | twitter 0.0058 | care 0.0050 | twitter 0.0081 |
| media 0.0041 | business 0.0050 | media 0.0048 | marketing 0.0075 |
| speech 0.0040 | blog 0.0041 | speech 0.0047 | business 0.0067 |
| social 0.0038 | top 0.0039 | social 0.0039 | video 0.0059 |
| white 0.0036 | obama 0.0037 | white 0.0038 | top 0.0056 |
| barack 0.0034 | great 0.0036 | reform 0.0037 | blog 0.0054 |
| house 0.0034 | video 0.0036 | house 0.0036 | great 0.0051 |

*Accuracy* Given user $u$, its label $s_u$ in the dataset and the assigned label $r_u$ obtained from the above methods, accuracy is defined as follows.

$$Accuracy = \frac{\sum_{u \in U} \delta(s_u, map(r_u))}{|U|}$$

where $|U|$ is the number of all the users and $\delta(x, y)$ is the delta function that is one if $x = y$ and is zero otherwise, and $map(r_u)$ is the permutation mapping function that maps the label $r_u$ of user $u$ to the corresponding label in the dataset. The best mapping between the labels can be found by Kuhn-Munkres algorithm [Kuhn 1955].

*Normalized Mutual Information* We denote $C$ as the user labels obtained from the dataset and $C'$ as the ones obtained from the above methods. The mutual information metric $MI(C, C')$ is defined as follows.

$$MI(C, C') = \sum_{c \in C, c' \in C'} p(c, c') \log \frac{p(c, c')}{p(c)p(c')}$$

where $p(c)$ is the probability that a user arbitrarily selected from the dataset has label $c$, and $p(c, c')$ is the joint probability that the arbitrarily selected document has label $c$ and is assigned with label $c'$. The normalized mutual information NMI is defined as follows.

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))} \tag{11}$$

where $H(C)$ is the entropy of $C$. Specifically, $NMI$ = 1 if $C$ and $C'$ are identical, and $NMI$ = 0 if $C$ and $C'$ are independent.

Table IX. Accuracy and Normalized Mutual Information by NormCut, SSNLDA and LCTA in DBLP dataset

| N | Accuracy(%) | | | NMI(%) | | |
|---|---|---|---|---|---|---|
| | NormCut | SSNLDA | LCTA | NormCut | SSNLDA | LCTA |
| 5 | 36.78 | 29.46 | **72.19** | 2.02 | 1.41 | **43.79** |
| 10 | 41.43 | 23.56 | **54.81** | 8.34 | 1.72 | **31.77** |
| 15 | 36.99 | 21.12 | **37.74** | 10.01 | 2.29 | **25.66** |
| 20 | 33.45 | 18.81 | **35.10** | 9.01 | 1.71 | **23.46** |

We show the result of accuracy and normalized mutual information in DBLP dataset in Table IX. From the table, we can see that LCTA performs the best among all the methods. The labels of the users in DBLP dataset mainly consider the coherence of topics. Since NormCut and SSNLDA do not consider the text information, both of them perform poor in the dataset. When the number of communities is large such as 15 and 20, NormCut performs relatively well in accuracy measure. The reason is that in the result of NormCut there is a small number of very large communities while in the result of other methods the clusters are of similar sizes. There are only four types of labels in DBLP dataset. If the result is dominant by one big cluster, the cluster can be mapped to one of the four labels, so the result will have a relatively large accuracy value. Therefore, accuracy may not be a good measure to compare the performance especially when the number of communities is large, so we use normalized mutual information as additional evaluation measure. In normalized mutual information both NormCut and SSNLDA perform poorly. Therefore we can see that our LCTA model performs better than those approaches processing community discovery and topic modeling separately.

### 4.4. Comparison with Topic Modeling Methods

In this section we compare our LCTA (Latent Community Topic Analysis) model with other topic modeling methods.

We compare the following methods in this section.

— PLSA [Hofmann 1999]: Probabilistic latent semantic analysis

— NetPLSA [Mei et al. 2008]: PLSA regularized with a harmonic regularizer based on a link graph structure

— LinkLDA [Erosheva et al. 2004]: A generative model of both text and links where words and links are generated according to the same latent topic space

— LCTA: Our Latent Community Topic Analysis model that integrates community discovery with topic modeling

The latent topics discovered by PLSA, NetPLSA and LinkLDA can also be used to calculate the clusters of users. Similarly accuracy and normalized mutual information (NMI) can be used to measure the clustering performance and topical coherence in DBLP dataset for comparing different methods.

Table X. Accuracy and Normalized Mutual Information in DBLP dataset by PLSA, NetPLSA, LinkLDA and LCTA

| N | Accuracy(%) | | | | NMI(%) | | | |
|---|---|---|---|---|---|---|---|---|
| | PLSA | NetPLSA | LinkLDA | LCTA | PLSA | NetPLSA | LinkLDA | LCTA |
| 5 | 67.55 | 68.09 | 69.33 | **72.19** | 41.29 | 42.49 | 40.50 | **43.79** |
| 10 | 44.54 | 45.96 | 44.05 | **54.81** | 28.71 | 29.05 | 26.86 | **31.77** |
| 15 | 33.39 | 34.23 | 32.48 | **37.74** | 22.36 | 23.09 | 21.72 | **25.66** |
| 20 | 25.89 | 26.42 | 26.67 | **35.10** | 20.06 | 21.24 | 18.73 | **23.46** |

The comparison result of both accuracy and normalized mutual information is listed in Table X. From the table we can see that our LCTA model performs the best among all the methods. Compared with PLSA, NetPLSA considers the link graph structure to regularize the topic modeling process, so it has a better performance than PLSA. Our LCTA model separates the concepts of topic and community and it performs better than LinkLDA in which both words and links are generated according to the same latent topic space.

Beside evaluating the topical coherence, we also compare the link structure coherence for different topic modeling methods. From the clusters of users by PLSA, NetPLSA and LinkLDA, we can calculate the normalized cut measure based on the partition of the graphs.

*Normalized Cut* Normalized cut (Ncut) is defined as $Ncut = \sum_{i=1}^{N} \frac{cut(C_i, U-C_i)}{assoc(C_i,U)}$, where $cut(A,B) = \sum_{u\in A, v\in B} n(u,v)$, $assoc(A,U) = \sum_{s\in A, t\in U} n(s,t)$ and $C_i$ is the i-th community. $cut(A,B)$ is the total weight of the edges that have been removed by disconnecting two parts $A$ and $B$. $assoc(A,U)$ is the total connection from nodes in $A$ to all the nodes in the graph.

Table XI. Normalized Cut in DBLP dataset by PLSA, NetPLSA, LinkLDA and LCTA

| N | PLSA | NetPLSA | LinkLDA | LCTA |
|---|---|---|---|---|
| 5 | 2.36 | 2.15 | 2.35 | **1.89** |
| 10 | 6.82 | 6.70 | 6.00 | **5.60** |
| 15 | 12.52 | 11.77 | 11.12 | **10.53** |
| 20 | 16.70 | 16.29 | 15.84 | **15.69** |

Table XII. Normalized Cut in Twitter dataset by PLSA, NetPLSA, LinkLDA and LCTA

| N | PLSA | NetPLSA | LinkLDA | LCTA |
|---|---|---|---|---|
| 5 | 7.85 | 6.87 | 7.10 | **6.11** |
| 10 | 17.80 | 16.25 | 16.78 | **16.24** |
| 15 | 27.80 | 26.29 | 26.76 | **25.91** |
| 20 | 37.72 | 35.84 | 36.23 | **35.47** |

Table XIII. Communities related to Hector Garcia-Molina, Rakesh Agrawal, Christos Faloutsos and Jiawei Han and their user distributions when the number of communities is 4

| Community 1 (Hector Garcia-Molina) | Community 2 (Rakesh Agrawal) |
|---|---|
| Hector Garcia-Molina 0.1395 | Rakesh Agrawal 0.1200 |
| Jennifer Widom 0.0455 | Ramakrishnan Srikant 0.0369 |
| Jeffrey D. Ullman 0.0425 | Surajit Chaudhuri 0.0340 |
| Rajeev Motwani 0.0325 | Michael J. Carey 0.0313 |
| Sharad Mehrotra 0.0213 | Raghu Ramakrishnan 0.0308 |
| Abraham Silberschatz 0.0190 | Hamid Pirahesh 0.0282 |
| Yannis Papakonstantinou 0.0186 | Jeffrey F. Naughton 0.0277 |
| Janet L. Wiener 0.0179 | David J. DeWitt 0.0262 |
| Serge Abiteboul 0.0163 | Jerry Kiernan 0.0190 |
| Wilburt Labio 0.0159 | Umeshwar Dayal 0.0175 |
| **Community 3 (Christos Faloutsos)** | **Community 4 (Jiawei Han)** |
| Christos Faloutsos 0.1513 | Jiawei Han 0.1719 |
| Andrew Tomkins 0.0300 | Jian Pei 0.0482 |
| Ravi Kumar 0.0298 | Wei Wang 0.0440 |
| Spiros Papadimitriou 0.0282 | Xifeng Yan 0.0302 |
| Jure Leskovec 0.0243 | Jeffrey Xu Yu 0.0239 |
| H. V. Jagadish 0.0236 | Ke Wang 0.0235 |
| Dimitris Papadias 0.0195 | Hongjun Lu 0.0222 |
| Jimeng Sun 0.0195 | Hong Cheng 0.0214 |
| Raymond T. Ng 0.0188 | Dong Xin 0.0209 |
| Agma J. M. Traina 0.0183 | Haixun Wang 0.0201 |

We list the result of DBLP dataset in Table XI and the one of Twitter dataset in Table XII. In both datasets, compared with other methods, LCTA performs better, which means that LCTA considers the link information relatively well and can discover more coherent communities.

### 4.5. Parameter Setting

In this section, we study the effect of parameter changes on the result. We have to set two parameters in our model, i.e., the number of communities and the number of topics.

To study the effect of the number of communities, we build a subset of DBLP dataset including all the co-authors of Hector Garcia-Molina, Rakesh Agrawal, Christos Faloutsos and Jiawei Han, which results in a graph of 494 users. We set the number of communities as 4 and 20 separately and compare the discovered communities. In Table XIII, we can see that if we set the number of communities as 4 there are four communities related to these four researchers. We increase the number of communities from 4 to 20 and show the selected communities related to Christos Faloutsos and Jiawei Han in Table XIV. Community 1 of Christos Faloutsos is about his community at Carnegie Mellon University and Community 2 is about his community along with Yahoo Research and Cornell University. Community 1 of Jiawei Han is about his community at University of Illinois at Urbana and Champaign including his students Xifeng Yan, Dong Xin, etc. Community 2 is another collaboration community of Jiawei Han. From this example, we can see that if we set the number of communities to a small value, we can have communities of coarse granularity. If we increase the number of communities, coarse communities will break down into the ones of fine granularity.

Table XIV. Selected communities and their user distributions when the number of communities is 20

| Selected communities related to Christos Faloutsos | |
|---|---|
| Community 1 | Community 2 |
| Christos Faloutsos 0.2322 | Ravi Kumar 0.1451 |
| Spiros Papadimitriou 0.0521 | Andrew Tomkins 0.1451 |
| Jimeng Sun 0.0426 | Christos Faloutsos 0.1043 |
| Dimitris Papadias 0.0412 | Jure Leskovec 0.0915 |
| Yufei Tao 0.0400 | Prabhakar Raghavan 0.0444 |
| Agma J. M. Traina 0.0386 | Carlos Guestrin 0.0349 |
| Caetano Traina Jr. 0.0297 | Andreas Krause 0.0349 |
| Hanghang Tong 0.0242 | Marko Grobelnik 0.0296 |
| Nick Roussopoulos 0.0220 | Jon M. Kleinberg 0.0296 |
| Jia-Yu Pan 0.0219 | Sridhar Rajagopalan 0.027 |
| Selected communities related to Jiawei Han | |
| Community 1 | Community 2 |
| Jiawei Han 0.2576 | Jiawei Han 0.1418 |
| Xifeng Yan 0.0950 | Wei Wang 0.1060 |
| Dong Xin 0.0654 | Anthony K. H. Tung 0.0919 |
| Xiaolei Li 0.0601 | Wen Jin 0.0792 |
| Hong Cheng 0.0594 | Haixun Wang 0.0767 |
| Hector Gonzalez 0.0271 | Hongjun Lu 0.0696 |
| Xiaoxin Yin 0.0237 | Martin Ester 0.0550 |
| Tianyi Wu 0.0224 | Jian Pei 0.0371 |
| Chen Chen 0.0195 | Nick Koudas 0.0326 |
| Feida Zhu 0.0180 | Jiong Yang 0.0301 |

Table XV. Selected topics in DBLP dataset when the number of topics is 20

| Selected topics related to database | | Selected topics related to machine learning | |
|---|---|---|---|
| Topic 1 | Topic 2 | Topic 1 | Topic 2 |
| query 0.0344 | queries 0.0385 | learning 0.0478 | image 0.0285 |
| database 0.0339 | temporal 0.0383 | bayesian 0.0226 | recognition 0.0268 |
| relational 0.0294 | spatial 0.0316 | networks 0.0189 | visual 0.0177 |
| queries 0.0284 | efficient 0.0274 | supervised 0.0185 | learning 0.0160 |
| databases 0.0229 | processing 0.0197 | analysis 0.0181 | motion 0.0144 |
| sql 0.0190 | objects 0.0163 | classification 0.0155 | shape 0.0129 |
| optimization 0.0185 | spatio 0.0153 | kernel 0.0124 | 3d 0.0117 |
| object 0.0150 | moving 0.0146 | models 0.0121 | vision 0.0111 |
| relations 0.0122 | indexing 0.0122 | fast 0.0103 | interactive 0.0107 |
| views 0.0117 | multidimensional 0.0119 | markov 0.0099 | fuzzy 0.0104 |

To study the effect of the number of topics, we vary the number of topics and compare the the results. If we set the number of topics as 4, from Table II, we can get the four topics including information retrieval, machine learning, data mining and database. If we increase the number of topics from 4 to 20, we can have the topics of fine granularity. In Table XV, we list several topics related to database and machine learning when the number of topics is 20. The first topic related to database is about relational database and query optimization, and the second is about spatial temporal database. The first topic related to machine learning is about bayesian networks and kernel methods, and the second is learning in computer vision.

## 5. RELATED WORK

In this section we discuss related work to our study including community discovery and topic modeling.

*Community discovery* Community discovery, a.k.a. group detection [Getoor and Diehl 2005], is to divide the network nodes into densely connected subgroups [Newman and Girvan 2004; Newman 2004b; Clauset et al. 2004; Palla et al. 2005; Leskovec et al. 2010], which is an important task in datasets including social networks [Parthasarathy et al. 2011], web graphs [Flake et al. 2000], biological networks [Girvan and Newman 2002], co-authorship networks [Newman 2004a], etc. Tang et al. [Tang and Liu 2010] provided a good overview of community discovery algorithms using network structures. Newman et al. [Newman and Girvan 2004] proposed an algorithm to remove edges from the network iteratively to split it into communities. The edges removed being identified using betweenness measures and the measures are recalculated after each removal. Palla et al. [Palla et al. 2005] analyzed the statistical features of overlapping communities to uncover the modular structure of complex systems. In [Ruan and Zhang 2007], Ruan et al. introduced an efficient spectral algorithm for modularity optimization to discover community structure. Nowicki et al. [Nowicki and Snijders 2001] proposed a statistical approach to a posteriori blockmodeling to partition the vertices of the graph into several latent classes where the probability distribution of the relation between two vertices depends only on the classes to which they belong. In [Zhang et al. 2007], Zhang et al. proposed an LDA-based hierarchical Bayesian algorithm called SSN-LDA, where communities are modeled as latent variables in the graphical models and defined as distributions over social actor space. In [Zhang et al. 2007], Zhang et al. used a Gaussian distribution with inverse-Wishart prior to model the arbitrary weights that are associated with the social interaction occurrences. Leskovec et al. [Leskovec et al. 2010] studied a range of network community detection methods originating from theoretical computer science, scientific computing, and statistical physics in order to compare them and to understand their relative performance and the systematic biases in the clusters they identify. All these studies focus on the link structure of the networks without considering the text information. In [Long et al. 2007], Long et al. proposed a probabilistic model for relational clustering under a large number of exponential family distributions and they also did not consider the topical coherence in the clustering process.

*Topic modeling* Statistical topic models can be considered as the probabilistic models for uncovering the underlying semantic structure of a document collection based on hierarchical Bayesian analysis of the text collection. Topic models, such as PLSA [Hofmann 1999] and LDA [Blei et al. 2003], use a multinomial word distribution to represent a semantic coherent topic and model the generation of the text collection with a mixture of such topics. Some studies extend topic modeling with networks. In [Mei et al. 2008], Mei et al. introduced a model called NetPLSA that regularizes a statistical topic model with a harmonic regularizer based on a graph structure in the data. In [Sun et al. 2009], Sun et al. defined a multivariate Markov Random Field for topic distribution random variables for each document to model the dependency relationships among documents over the network structure. In these studies the links in the graph are not modeled in a generative process. Zhou et al. [Zhou et al. 2006] proposed a generative probabilistic model to discover semantic community in social networks, but they used text information only without considering link structure. There are several studies on generative topic models based on text and links including Author-Topic model [Steyvers et al. 2004; Rosen-Zvi et al. 2004], Author-Recipient-Topic model [McCallum et al. 2005; McCallum et al. 2007; Pathak et al.

2008], Group-Topic model [Wang et al. 2005], Link-PLSA-LDA [Nallapati and Cohen 2008], Block-LDA [Balasubramanyan and Cohen 2011], Topics-on-Participations model [Zheng et al. 2010; 2011]. Cohn et al. [Cohn and Hofmann 2000] proposed a joint probabilistic term-citation model where the generation of each link in a document is a multinomial sampling of the document. Following [Cohn and Hofmann 2000], Erosheva et al. [Erosheva et al. 2004] used a mixed membership model for words and references but treated the membership scores as random Dirichlet realizations. In [Cohn and Hofmann 2000; Erosheva et al. 2004] both text and links are from the same topic-specific space, so they cannot capture the topical coherence in the community level as our model does. Liu et al. [Liu et al. 2009] proposed a model called Topic-Link LDA where the membership of authors is modeled with a mixture model and whether a link exists between two documents follows a binomial distribution parameterized by the similarity between topic mixtures and community mixtures as well as a random factor. In [Wang and Blei 2011] Wang et al. combined the merits of collaborative filtering and probabilistic topic modeling whereas in our model we integrate community discovery with topic modeling. Deng et al. [Deng et al. 2011] proposed a topic model with biased propagation algorithm to incorporate heterogeneous information network with topic modeling. However, it is not related to community discovery, and the discovered topics are not community-based either. In our model community and topic are different concepts. One community can correspond to multiple topics and multiple communities can share the same topic. The interaction of communities and topics provides flexibility in the community discovery process. We also compare our method with topic modeling methods in our experiment.

## 6. CONCLUSION AND FUTURE WORK

With the development of social media a lot of user-generated content is available with user networks. The user graphs extended with text information on the nodes form text-associated graphs. In this paper we study the problem of latent community topic analysis in text-associated graphs and propose a model called LCTA to incorporate community discovery into topic modeling. We handle topic modeling and community discovery in the same framework to guarantee the topical coherence in the communities. We perform extensive experiments on two real datasets and show that our model outperforms other methods.

Our work opens up several interesting future directions. First, the communities in our LCTA model are of the same level but the communities in real world may have hierarchical structure. It is interesting to extend our framework to hierarchical community discovery scenarios by bottom-up or top-down strategy. Second, the user-generated content in social media sites includes not only text data but also other rich information such as pictures, time, and spatial information. It is interesting to integrate those heterogeneous information together in the framework.

## ACKNOWLEDGMENTS

The work was supported in part by U.S. National Science Foundation grants CCF-0905014, CNS-0931975, and IIS-09-05215, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), and U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government or the Army Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## REFERENCES


BALASUBRAMANYAN, R. AND COHEN, W. W. 2011. Block-lda: Jointly modeling entity-annotated text and entity-entity links. In *Proc. of 2011 SIAM Int. Conf. on Data Mining (SDM'11)*. 450–461.

BLEI, D. M. 2011. Introduction to probabilistic topic models. In *Communications of the ACM*.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

CAI, D., MEI, Q., HAN, J., AND ZHAI, C. 2008. Modeling hidden topics on document manifold. In *Proc. of 2008 ACM Conf. on Information and Knowledge Management (CIKM'08)*. 911–920.

CLAUSET, A., NEWMAN, M. E. J., AND MOORE, C. 2004. Finding community structure in very large networks. *Physical Review E 70*, 066111.

COHN, D. A. AND HOFMANN, T. 2000. The missing link - a probabilistic model of document content and hypertext connectivity. In *Proc. of 2000 Neural Info. Processing Systems Conf. (NIPS'00)*. 430–436.

DENG, H., HAN, J., ZHAO, B., YU, Y., AND LIN, C. X. 2011. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proc. of 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*. 1271–1279.

EROSHEVA, E., FIENBERG, S., AND LAFFERTY, J. 2004. Mixed-membership models of scientific publications. *Proc. of the National Academy of Sciences 101*, 5220–5227.

FLAKE, G. W., LAWRENCE, S., AND GILES, C. L. 2000. Efficient identification of web communities. In *Proc. of 2000 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'00)*. 150–160.

GAO, J., FAN, W., SUN, Y., AND HAN, J. 2009. Heterogeneous source consensus learning via decision propagation and negotiation. In *Proc. of 2009 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09)*. 339–348.

GETOOR, L. AND DIEHL, C. P. 2005. Link mining: a survey. *SIGKDD Explorations 7,* 2, 3–12.

GIRVAN, M. AND NEWMAN, M. E. 2002. Community structure in social and biological networks. *Proc. of the National Academy of Sciences 99,* 12, 7821–7826.

HOFMANN, T. 1999. Probabilistic latent semantic indexing. In *Proc. of 1999 Int. ACM SIGIR Conf. on Research & Development in Information Retrieval (SIGIR'99)*. 50–57.

KUHN, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly 2*, 83–97.

KWAK, H., LEE, C., PARK, H., AND MOON, S. B. 2010. What is twitter, a social network or a news media? In *Proc. of 2011 Int. World Wide Web Conf. (WWW'11)*. 591–600.

LESKOVEC, J., LANG, K. J., AND MAHONEY, M. W. 2010. Empirical comparison of algorithms for network community detection. In *Proc. of 2010 Int. World Wide Web Conf. (WWW'10)*. 631–640.

LIN, J. AND DYER, C. 2010. *Data-Intensive Text Processing with MapReduce*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

LIU, Y., NICULESCU-MIZIL, A., AND GRYC, W. 2009. Topic-link lda: joint models of topic and author community. In *Proc. of 2009 Int. Conf. on Machine Learning (ICML'09)*. 84.

LONG, B., ZHANG, Z. M., AND YU, P. S. 2007. A probabilistic framework for relational clustering. In *Proc. of 2007 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'07)*. 470–479.

MCCALLUM, A., CORRADA-EMMANUEL, A., AND WANG, X. 2005. Topic and role discovery in social networks. In *Proc. of 2005 Int. Joint Conf. on Artificial Intelligence (IJCAI'05)*. 786–791.

MCCALLUM, A., WANG, X., AND CORRADA-EMMANUEL, A. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research 30*, 249–272.

MEI, Q., CAI, D., ZHANG, D., AND ZHAI, C. 2008. Topic modeling with network regularization. In *Proc. of 2008 Int. World Wide Web Conf. (WWW'08)*. 101–110.

MEI, Q., LIU, C., SU, H., AND ZHAI, C. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proc. of 2006 Int. World Wide Web Conf. (WWW'06)*. 533–542.

NALLAPATI, R. AND COHEN, W. W. 2008. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *Proc. of 2008 Int. AAAI Conference on Weblogs and Social Media (ICWSM'08)*.

NEWMAN, M. E. J. 2004a. Coauthorship networks and patterns of scientific collaboration. *Proc. of the National Academy of Sciences*, 5200–5205.

NEWMAN, M. E. J. 2004b. Fast algorithm for detecting community structure in networks. *Physical Review E 69*, 066133.

NEWMAN, M. E. J. AND GIRVAN, M. 2004. Finding and evaluating community structure in networks. *Physical Review E 69*, 026113.

NOWICKI, K. AND SNIJDERS, T. A. B. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96,* 455, 1077–1087.

PALLA, G., DERENYI, I., FARKAS, I., AND VICSEK, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*, 814.

PARTHASARATHY, S., RUAN, Y., AND SATULURI, V. 2011. Community discovery in social networks: Applications, methods and emerging trends. In *Social Network Data Analytics*. Springer US, 79–113.

PATHAK, N., DELONG, C., BANERJEE, A., AND ERICKSON, K. 2008. Social topic models for community extraction. In *Proc. of 2008 SNA-KDD Workshop on Social Network Mining and Analysis (SNA-KDD'08)*.

ROSEN-ZVI, M., GRIFFITHS, T. L., STEYVERS, M., AND SMYTH, P. 2004. The author-topic model for authors and documents. In *Proc. of 2004 Int. Conf. on Uncertainty in Artificial Intelligence (UAI'04)*. 487–494.

RUAN, J. AND ZHANG, W. 2007. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In *Proc. of 2007 Int. Conf. on Data Mining (ICDM'07)*. 643–648.

SHI, J. AND MALIK, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 22,* 8, 888–905.

STEYVERS, M., SMYTH, P., ROSEN-ZVI, M., AND GRIFFITHS, T. L. 2004. Probabilistic author-topic models for information discovery. In *Proc. of 2004 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'04)*. 306–315.

SUN, Y., HAN, J., GAO, J., AND YU, Y. 2009. itopicmodel: Information network-integrated topic modeling. In *Proc. of 2009 Int. Conf. on Data Mining (ICDM'09)*. 493–502.

TANG, L. AND LIU, H. 2010. *Community Detection and Mining in Social Media*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

WANG, C. AND BLEI, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proc. of 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*. 448–456.

WANG, X., MOHANTY, N., AND MCCALLUM, A. 2005. Group and topic discovery from relations and their attributes. In *Proc. of 2005 Neural Info. Processing Systems Conf. (NIPS'05)*.

ZHAI, C., VELIVELLI, A., AND YU, B. 2004. A cross-collection mixture model for comparative text mining. In *Proc. of 2004 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'04)*. 743–748.

ZHANG, H., GILES, C. L., FOLEY, H. C., AND YEN, J. 2007. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *Proc. of 2007 AAAI Conf. on Artificial Intelligence (AAAI'07)*. 663–668.

ZHANG, H., QIU, B., GILES, C. L., FOLEY, H. C., AND YEN, J. 2007. An lda-based community structure discovery approach for large-scale social networks. In *Proc. of 2007 IEEE Conf. on Intelligence and Security Informatics (ISI'07)*. 200–207.

ZHENG, G., GUO, J., YANG, L., XU, S., BAO, S., SU, Z., HAN, D., AND YU, Y. 2010. A topical link model for community discovery in textual interaction graph. In *Proc. of 2010 ACM Conf. on Information and Knowledge Management (CIKM'10)*. 1613–1616.

ZHENG, G., GUO, J., YANG, L., XU, S., BAO, S., SU, Z., HAN, D., AND YU, Y. 2011. Mining topics on paritcipations for community discovery. In *Proc. of 2011 Int. ACM SIGIR Conf. on Research & Development in Information Retrieval (SIGIR'11)*. 445–454.

ZHOU, D., MANAVOGLU, E., LI, J., GILES, C. L., AND ZHA, H. 2006. Probabilistic models for discovering e-communities. In *Proc. of 2006 Int. World Wide Web Conf. (WWW'06)*. 173–182.