Volume 6, Number 1, 2012

ISSN 1751-1577

ELSEVIER

Journal of
INFORMETRICS
An International Journal

Editor-in-Chief: **Leo Egghe**

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Journal of Informetrics 6 (2012) 237–253

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Adding community and dynamic to topic models

Daifeng Li[a], Ying Ding[c], Xin Shuai[b,*], Johan Bollen[b], Jie Tang[a], Shanshan Chen[c], Jiayi Zhu[b], Guilherme Rocha[d]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[b] School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[c] School of Library and Information Science, Indiana University, Bloomington, IN, USA
[d] Department of Statistics, Indiana University, Bloomington, IN, USA

ARTICLE INFO

Article history:
Received 21 April 2011
Received in revised form 4 October 2011
Accepted 7 November 2011

Keywords:
Social network
Semantic community
Topic mining
Dynamic

ABSTRACT

The detection of communities in large social networks is receiving increasing attention in a variety of research areas. Most existing community detection approaches focus on the topology of social connections (e.g., coauthor, citation, and social conversation) without considering their topic and dynamic features. In this paper, we propose two models to detect communities by considering both topic and dynamic features. First, the Community Topic Model (CTM) can identify communities sharing similar topics. Second, the Dynamic CTM (DCTM) can capture the dynamic features of communities and topics based on the Bernoulli distribution that leverages the temporal continuity between consecutive timestamps. Both models were tested on two datasets: ArnetMiner and Twitter. Experiments show that communities with similar topics can be detected and the co-evolution of communities and topics can be observed by these two models, which allow us to better understand the dynamic features of social networks and make improved personalized recommendations.

1. Introduction

The study of social networks has enabled scientists to better understand social communication patterns and interpret social principles. Researchers have found that most real world networks, in contrast to random networks, exhibit three common properties: the small world property, power-law distributions, and community structure with relatively high clustering coefficient (Erdos & Renyi, 1959; Girvan & Newman, 2002; Milgram, 1967; Newman, 2001). In fact, evidence of communities has been detected in a range of different domains and applications (Leskovec, Lang, & Mahoney, 2010). Communities detected within a social network might correspond to a variety of social groupings affected by the heterogeneity of users as well as their interactions. The analysis of communities is therefore crucial to establish a better understanding and utilization of social networks (Flake, Tarjan, & Tsioutsiouliklis, 2003). However, most of the existing work has focused on the structural properties of communities and neglects other important aspects such as their topic features. In addition, the structural properties and topic aspects of communities may interact with each other. Common interests may drive the formation of communities, and in turn community structure may reinforce common interests. Few studies have systematically and quantitatively addressed the interaction between the structural and topic properties of communities (Ding, 2011).

Furthermore, social networks and their communities may change over time. Any effort to understand the formation of communities and their topical features needs to include the time dimension. Previous studies have used state space models

[*] Corresponding author at: School of Informatics and Computing, Indiana University, 919 E. 10th Street, Rm 401, Bloomington, IN 47408, USA.
E-mail address: xshuai@indiana.edu (X. Shuai).

1751-1577/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2011.11.004

on the natural parameters of multinomial distributions to analyze the time evolution of topics, or developed the continuous time dynamic model to mine the latent topics through a sequential collection of documents (Blei & Lafferty, 2006; Griffiths & Steyvers, 2004; Iwata, Yamada, Sakurai, & Ueda, 2010; Wang, Blei, & Heckerman, 2008). Generally, these studies have applied a set of approaches to approximate posterior inference over the latent topics. However, none considered the community features of the network actors involved in their datasets that might reveal some hidden explanation for topical evolution.

To address those challenging problems for detecting communities by considering their topic features, we propose the Community Topic Model (CTM). To further capture the dynamic features of community evolution, we propose the Dynamic Community Topic Model (DCTM) by extending CTM with the time variables. Both CTM and DCTM were applied to two large-scale datasets: Arnetminer (Scholarly publications in the area of Computer Science) and Twitter. The experiments show that both models can capture the topic features and dynamics changes of communities.

This paper is organized as follows: Section 2 defines the problem. Section 3 provides an overview of related work and Section 4 discusses the methods, datasets as well as the proposed CTM and DCTM. Section 5 analyzes the results by applying the proposed models to two large datasets. Section 6 discusses the findings and Section 7 concludes the study.

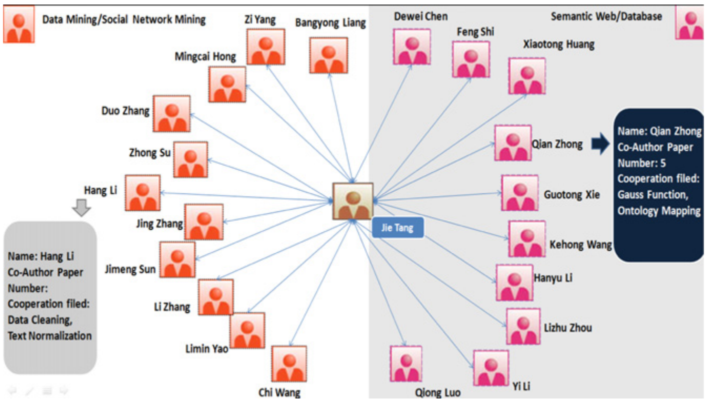
2. Problem definition

In a social network, actors may have different topic interests and therefore can be divided into different communities according to their topic distributions. For example, in Fig. 1, the author has different research focuses: the Semantic Web and Text Mining, which means that he can belong to two communities with different topics. Most existing community detection methods focus on the topological structures of networks and ignore actors' topic interests.

In order to detect communities from the topic level, the proposed algorithm should assign appropriate actors for each community based on matching topic interests. Unlike Author–Topic model (Rosen-zvi, Griffiths, Steyvers, & Smyth, 2008) that assigns authors to different topics based on the authors' topic distributions, the proposed CTM assigns actors to different communities based on the similarity between authors' topic distributions and community's topic distributions. For example, researchers, who work in both areas of biology and the Semantic Web can be viewed as having similar topic distributions. Therefore, these researchers can be grouped into one community. It is hard to find the similar topic distributions of a group of authors, because the relationship between a group of authors and their topic distributions are latent variables. There are studies emphasizing how these two variables jointly affect the formation of links in the document graph (Zhu, Yu, Chi, & Gong, 2007). However, some questions are left unanswered, such as whether the structure of communities has an influence on the distribution of topics and how topic distribution determines the features of a community.

Additionally, an examination of changes over time is needed in order to discover the dynamic relationship between communities and topics. Traditional methods treat different timestamps independently and ignore the temporal continuity between consecutive timestamps (He et al., 2009; Li et al., 2010). These studies have two problems. The first problem is how to determine the corresponding relationship between latent variables from different timestamps (Griffiths & Steyvers, 2004; He et al., 2009). For instance, for a certain community at time t, it is hard to know which community it was derived from time $t-1$. The previous method requires calculating the similarity between the current community in time t and all communities in time $t-1$ in order to figure out the temporal inheritance. The community–topic distribution and the topic–word distribution were needed for each calculation, which can be computationally expensive (Li et al., 2010). The second problem is that the temporal correlation between consecutive timestamps was not considered. For example, an author's previous research interests may influence his current interests. The proposed DCTM can simulate the changes of actors' interests at different time periods, and observe the evolution of communities and topics along the time.

Taking the example below, an actor in a social network can be defined as actor $=(a, \{z_1, t_1\}, \{z_2, t_2\}, \{z_3, t_3\}, \ldots)$, where a means the actor, and $\{z_i, t_i\}$ is the tuple which represents actor a created an action z_i at time point t_i. We propose the following function to define communities: $c(a_i, a_j, c_k, t) = f(S_{a_i}, S_{a_j}, S_{c_k}, t)$. Here, $c(a_i, a_j, c_k, t)$ represents the decision to put



Name: Hang Li
Co-Author Paper Number:
Cooperation filed: Data Cleaning, Text Normalization

Name: Qian Zhong
Co-Author Paper Number: 5
Cooperation filed: Gauss Function, Ontology Mapping

Fig. 1. Example that one author may have two topics.

Table 1
Notation table.

Notation	Meaning
d	Document
w	Word
x	Author
z	Topic
r	Publication venue (e.g., conference)
c	Community
N_d	The number of words in the current document d
N_D	The number of words in the entire collection of documents
a_d	The set of co-authors in paper d
α	Hyperparameter for generating θ from Dirichlet Distribution
λ	Hyperparameter for generating χ from Dirichlet Distribution
β	Hyperparameter for generating φ from Dirichlet Distribution
μ	Hyperparameter for generating ψ from Dirichlet Distribution
χ	A multinomial distribution of authors over communities
θ	A multinomial distribution of communities over topics
φ	A multinomial distribution of topics over words
ψ	A multinomial distribution of topics over publication venues
D	Collection of documents
A	Collection of authors
T	Collection of topics
R	Collection of conferences

two authors a_i and a_j into a community c_k at time t, which depends on the formula $f(S_{a_i}, S_{a_j}, S_{c_k})$, where $S_{a_i}, S_{a_j}, S_{c_k}$ means the topic distribution of a_i, a_j, c_k at time t. Table 1 summarizes the mathematical notation used in this paper.

3. Related work

3.1. Community detection

Researchers have used a number of methods to detect communities within networks. Two widely used approaches are those based on centrality and graph partitioning. Girvan and Newman (2002) used betweeness centrality to examine the community structure in large networks. The original algorithm was improved upon by Clauset, Moore, and Newman (2008), who reduced the complexity from $O(m^2n)$ to $O(md \log n)$ (where d is the depth of the dendrogram of the community structure). This algorithm has been tested empirically and validated as an appropriate model for community detection (Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004). Two standard examples of the graph partitioning approach are the local spectral partitioning algorithm (Andersen, Chung, & Lang, 2008) and the flow-based Metis_MQI algorithm (Flake et al., 2003). These approaches were compared to the Girvan–Newman algorithm by Leskovec et al. (2010). In applying all of these algorithms against the same large-scale dataset, Leskovec et al. (2010) found that the algorithms produced similar results and identified equally compact clusters at all scale sizes. However, none of these algorithms have taken into consideration the topic feature of communities.

3.2. Topic modeling

Since the introduction of the LDA model (Blei, Ng, & Jordan, 2003), various extended LDA models have been used for automatic topic extraction from large-scale corpora. Rosen-zvi et al. (2008) introduced the Author–Topic model, which extended LDA to include authorship as a latent variable. Based on the Author–Topic model, Tang, Jin, and Zhang (2008) further extended the LDA and Author–Topic model and proposed the Author–Conference–Topic (ACT) model, which is a unified topic model for simultaneously modeling different types of information in academic networks. Nallapati and Cohen (2008) proposed a Link-PLSA-LDA model as a scalable LDA-type model for topic modeling and link prediction. Later, Si and Sun (2009) proposed a tag-LDA model, which extended the LDA model by adding a tag variable, and applied it to social tagging systems. The link structure of networks has served as an additional area for network research. Chang and Blei (2009) introduced the relational topic model (RTM) to model the link between documents as a binary random variable conditioned on their contents. Although research has been done in both areas of community detection and topic analysis, very few researchers have sought to combine the two. One notable exception is the work of Zhou, Manavoglu, Li, Giles, and Zhai (2006), who used topic model for semantic community discovery in social network analysis. The other is the work of Liu, Niculescu-Mizil, and Gryc (2009) who examined topic and author communities for a set of blog posts and citation data through jointly modeling underlying topics, author community, and link formation in one unified model. However, it was done synchronically, rather than diachronically. Therefore, it did not provide an evaluation of how the model functions in examining changes in topics over time.

As discussed above, studies on community detection have not taken other aspects of community profile into consideration, while research on topic modeling largely neglects potential relationships between topics and community structure. In this paper, we propose a different approach to address this question, by integrating dynamics and communities into the topic modeling algorithms.

4. Methods

In this paper, CTM (Community Topic Model) and DCTM (Dynamic Community Topic Model) were proposed to capture the semantic relationships among communities and topics as well as their changes over time. Two datasets, Arnetminer and Twitter, were used to test these two models.

4.1. Datasets

4.1.1. Arnetminer dataset

ArnetMiner (http://www.arnetminer.org) is an academic search system developed by the Tshinghua University (Tang, Zhang, et al., 2008). The Arnetminer dataset covers the major publications in the area of computer science. It was collected by using a unified automatic extraction approach on researcher's profile pages from the Web and other online digital libraries. Currently, this dataset contains 629,814 publications, 12,609 conferences, and 595,740 authors covering the period of 2000–2010 (Tang, Zhang, et al., 2008). Each publication has the information about abstract, authors, year, venue, and title. The abstracts and titles were pre-processed using a stemming algorithm and a stop word list.

4.1.2. Twitter dataset

In Twitter, the hashtag is a special tag starting with '#', like #teamlakers or #science. The hashtag is used to group tweets with similar topics, which is functionally similar to the publication venue of an academic paper. Tweets were crawled via Twitter streaming API from July 9, 2010 to September 9, 2010, and only those tweets with hashtag were selected. The 30 most frequent hashtags from each week were selected to represent the hot topics in Twitter during that time period. Furthermore, 20,000 tweets containing these 30 hashtags were selected randomly for each week and the retweets were removed to prevent repeated information. The original person who posted the tweet as well as the mentioned usernames in the tweet are the authors of the tweet. After the preprocessing, the Twitter dataset contains 152,768 tweets, 104,571 authors, and 158 hashtags.

4.2. Background knowledge

To better understand the algorithm of CTM, we will introduce the concepts of Dirichlet distribution, *sKL*, F1-measure and Gibbs sampling in this section.

(1) Dirichlet Distribution: Dirichlet Distribution is a family of continuous multivariate probability distribution, which is used to denote the probability of a probability event. There are two main reasons for us to apply it to LDA (Latent Dirichlet Allocation): first, it is the conjugate prior of the categorical distribution and multinomial distribution, which can help us to solve the model by applying Gibbs sampling algorithm; second, it can provide initial parameters estimation, which can train LDA model to learn training data and analyze new data.

(2) Gibbs sampling: Gibbs sampling is an efficient algorithm for solving MCMC (Monte Carlo–Markov Chain) problem. The process of learning topic distribution of training dataset in LDA can be seen as a MCMC process. Gibbs sampling can help the process of learning become more and more accurate after many step of iterations.

(3) *sKL*: *sKL* is used to compute the similarity between two variables represented by feature vectors. Assuming we have two authors x_1, x_2 with interesting distribution over ten topics, then the *sKL* value between the two authors can be seen as below:

$$sKL(x_1, x_2) = \sum_{i=1}^{10} \left[xz_{1i} \times \log \frac{xz_{1i}}{xz_{2i}} + xz_{2i} \times \log \frac{xz_{2i}}{xz_{1i}} \right] \tag{1}$$

we can find that the lower the *sKL* is for two variables, the more similar the two variables are.

(4) F1-measure: we often use precision and recall to evaluate the performance of an algorithm's prediction power. Precision is the percentage of predictions that are correct while recall is the percentage of total number of correct predictions that are achieved. F1-measure integrates precision and recall to give a comprehensive evaluation for the performance of target model.

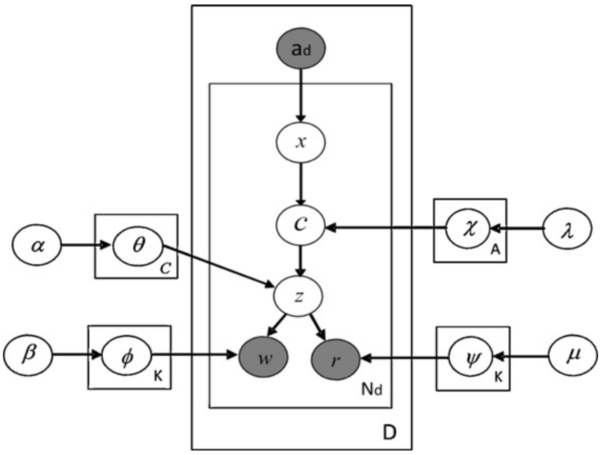


Fig. 2. Community Topic Model (CTM).

4.3. CTM (Community Topic Model)

The essential idea of CTM is to detect communities based on topic distributions over all authors and cluster authors with similar topic distributions together in one community. The community detection is achieved through a statistical learning process, during which the assignment of an author to a certain community (also including other types of assignments, like assigning topic to author, conference to topic, etc.) is implemented by sampling from several continuously updated and mutually related probability distributions.

The graphical representation of the learning process is shown in Fig. 2, which can be explained by the following example. A group of authors, a_d, collaborate on a document/paper d. For each author x in a_d, x first selects a community c from the author–community distribution χ; then select a topic z under community c from the topic-community distribution θ; and finally select a word w under topic z from the topic–word distribution ϕ, and a conference r related to z from the topic-conference distribution ψ. The three shaded nodes, a_d, w and r, are all observable.

Gibbs sampling is used to estimate 4 parameters θ, χ, ψ and ϕ, whose initial values (i.e. prior probabilities) are determined by another 4 hyperparameters: λ, α, β and μ and their empirical values are given by $\lambda = 50/C, \alpha = 50/T, \beta = 0.01$ and $\mu = 0.1$ (Lu et al., 2010; Rosen-zvi et al., 2008; Tang, Zhang, et al., 2008; Tang, Jin, et al., 2008). The final value of θ, χ, ψ and ϕ are obtained after 1000 iterations of sampling and estimation.

In each iteration, CTM assigns a community $c \in C$ and a topic $z \in T$, to each author $x \in A$ and each word $w \in V$, appeared in every document $d \in D$. For every possible assignment $(c, z) \in C \times T$, the following probability is calculated:

$$P_{c,z}(z^i = z, c^i = c, x^i = x | w^i = w, r^i = r, z^{i-1}, x^{i-1}, w^{i-1}, a_d) \propto \chi^{i-1} \cdot \theta^{i-1} \cdot \phi^{i-1} \cdot \psi^{i-1} \tag{2}$$

where i and $i-1$ denotes the corresponding values of the current and previous step of iteration; $P_{c,z}$ denotes the probability that (c,z) is assigned to x given the previous estimated results and current observation. Based on the multinomial distribution $\{P_{c,z}, (c, z) \in C \times T\}$, a community c and a topic z will be randomly sampled and assigned to author x and word w. In addition, topic z is assigned to conference r without sampling but from direct observation. After all assignments are done in the ith iteration, all probability matrices are updated as:

$$\chi_{x,c} = \frac{n_{x,c} + \lambda}{\sum_{c' \in C}(n_{x,c'} + \lambda)}, \qquad \theta_{c,z} = \frac{n_{c,z} + \alpha}{\sum_{z' \in T}(n_{c,z'} + \alpha)}, \qquad \phi_{z,w} = \frac{n_{z,w} + \beta}{\sum_{w' \in V}(n_{z,w'} + \beta)}, \qquad \psi_{z,r} = \frac{n_{z,r} + \mu}{\sum_{r' \in R}(n_{z,r} + \mu)} \tag{3}$$

$\chi_{x,c}$ denotes the entry of author x and community c in matrix χ, and $n_{x,c}$ denotes the number of times author x is assigned to community c at the current iteration. The similar denotation applies to $\theta_{c,z}$, $\phi_{z,w}$, and $\psi_{z,r}$.

For each iteration, parameters estimated from previous iteration are utilized to make re-assignment for all the authors. The final probability distribution matrices, i.e. θ, χ, ψ and ϕ, will be very close to the actual value when the iteration is done. In other words, the assignment of authors to communities will be sufficiently accurate in the end. In the process of statistical learning, the assignments of topics to communities are determined by all the authors' topic assignments, and authors with similar topic distribution are most likely to be assigned to the same community as the CTM estimation becomes more and more accurate. Besides, the rank of an author in a community (the probability of assigning the author to the community) is determined by his/her interest in the most popular topics (topics with high probabilities of being assigned to the community) in that community. For instance, if an author is interested in several topics and frequently writes papers to those topics, he/she is very likely to be assigned to the communities in which those topics rank high, and become a highly ranked author in that community. The algorithm in mathematic language is shown as the following:

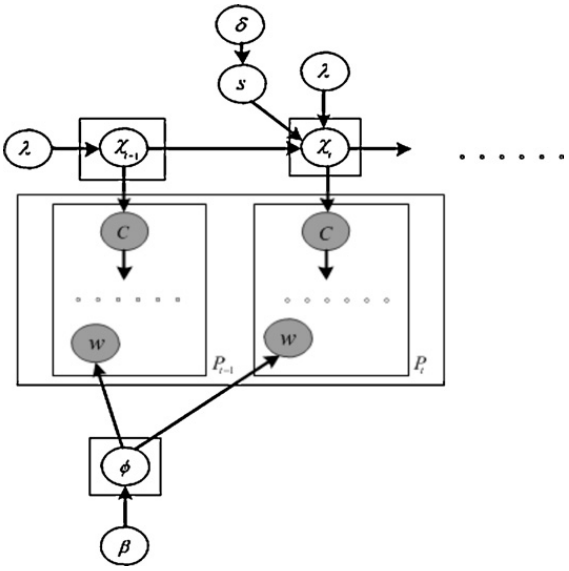


Fig. 3. Dynamic Community Topic Model (DCTM).

```
For $i = 1{:}1000$ iterations:
  For each document $d$
  For each author $x$ in document $d$:
    For each word $w$ in document $d$:
      Compute Eq. (2) and sample a community $c$ and topic $z$ to the current author $x$ and word $w$; also assign topic $z$ to conference $r$.
      Update $\chi$, $\phi$, $\psi$ and $\theta$ according to Eq. (3);
    End for word $w$;
  End for author $x$;
  End for document $d$;
End for iterations.
Output final $\chi$, $\phi$, $\psi$ and $\theta$.
```

4.4. DCTM (Dynamic Community Topic Model)

To model the evolution of a community, we assume that the distributions of communities are based on a Bernoulli trial. When time goes from one time slice to another, we flip a coin for each author. If the coin lands as a head, the previous community distribution will be kept. Otherwise, a new distribution will be sampled for that author. To determine authors' current interest, a switch variable s is introduced. The value of $s(s \in \{0, 1\})$ is sampled based on a Bernoulli distribution δ. When the sampled value of s equals 1, author's current interest is determined by his status in the last time period; when the sampled value of s equals 0, author's current interest is not influenced by his previous status but his current status.

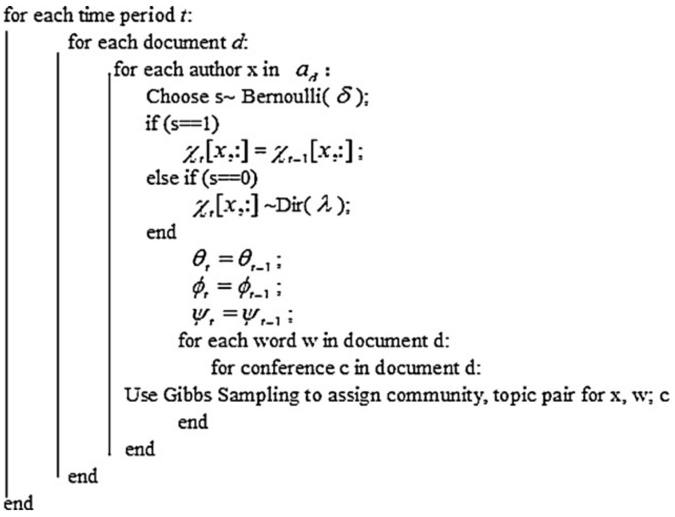
In Fig. 3, the results from previous time point $t - 1$ are used as prior knowledge to train the current training dataset at the time point t, and the Bernoulli trial is applied to simulate the changes of authors' interests. The dynamic model assigns a unique id for each author, community, topic, word and conference at the first time period, and passes these to the next time period after the iteration of the first time period is finished. Therefore, all communities and topics from different time slices can be consistently tracked. The pseudo-code of DCTM can be seen in Fig. 4.

5. Result

5.1. Result analysis from the static perspective

For the Arnetminer dataset, the whole time span was divided into three periods: 2000–2003, 2004–2007, and 2008–2010. In each time period, CTM was used to calculate the topic distribution of author, community and conference. The probability distribution of author for a given community was used to assign authors to different communities. 20 communities and 30 topics were extracted using the CTM. Authors in each community detected by CTM have similar topic distributions.

CTM can calculate the author community distribution and community topic distribution, while other existing models could not. These distributions provide enriched information to analyze the relationships among author, community and topic. Fig. 5a displays the author community distributions for the selected 10,000 authors during the period of 2008–2010. This can be explained as the authors' community preference. The value in the y-axis indicates the probability of an author choosing a community. Some authors have very high preference for certain communities such as Community 1 (computer system, network), Community 2 (intelligence system, parallel and distributed systems, semantic web, neural, wireless network, fuzzy), and Community 15 (image recognition, knowledge management, mathematics, machine learning, user interface and



```
for each time period t:
    for each document d:
        for each author x in $a_d$ :
            Choose s~ Bernoulli( $\delta$ );
            if (s==1)
                $\chi_t[x,:] = \chi_{t-1}[x,:]$ ;
            else if (s==0)
                $\chi_t[x,:]$ ~Dir( $\lambda$ );
            end
                $\theta_t = \theta_{t-1}$ ;
                $\phi_t = \phi_{t-1}$ ;
                $\psi_t = \psi_{t-1}$ ;
                for each word w in document d:
                    for conference c in document d:
            Use Gibbs Sampling to assign community, topic pair for x, w; c
                end
            end
        end
end
```

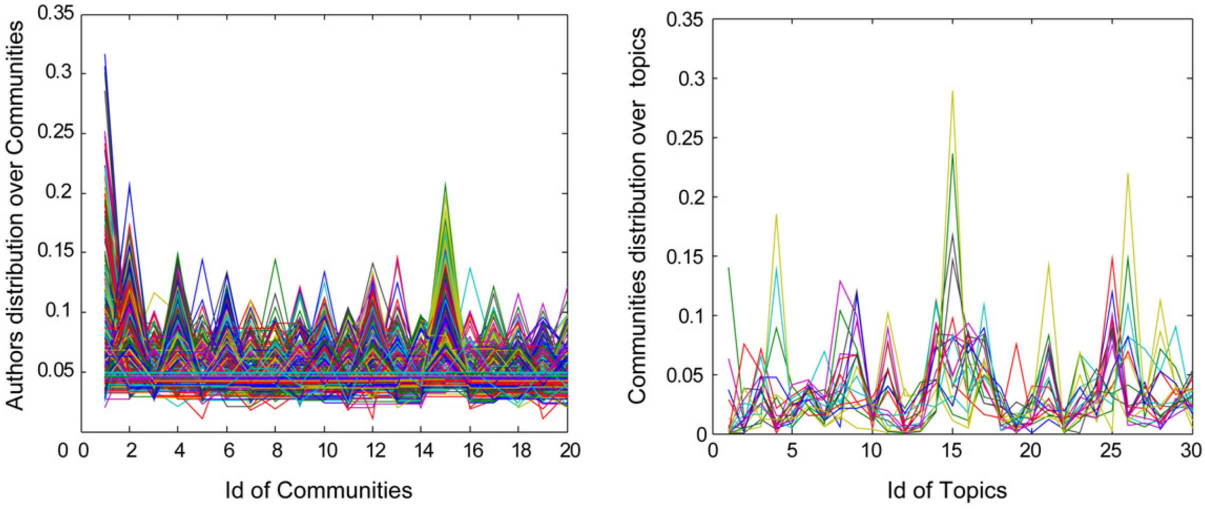
Fig. 4. Algorithm description for DCTM.

collaboration system). Fig. 5b shows the community topic distribution. For each community, the probability of a given topic (i.e. the value in y-axis) indicates the significance of this topic in that community. Some communities have strong preferences on Topic 4 (manufacturing optimization), Topic 15 (parallel and distributed systems), Topic 21 (embedded systems), and Topic 26 (knowledge management).

The value of y-axis in Fig. 6 represents the mean probability of all authors choosing a community, which can be inferred as the popularity of a community. Fig. 6 shows the popularity of communities during the period of 2008–2010. The range of popularity for all communities is between 0.049 and 0.053, indicating that the popularity differences among all communities are small. Community 15 is the most popular community in 2008–2010.

Fig. 7 shows the topic distribution for Community 15. Topic 16 and 26 are the most popular topics in Community 15, followed by topic 17, 25 and 28. Table 2 illustrates the top words, conferences for above popular topics and the top authors for the Community 15. Topics in Community 15 are diverse including image recognition, software development, wireless network, information management and mathematic algorithms. The listed top journals and conferences are consistent with the content of the topics, for example, PAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence) in Topic 16, SSEN (ACM SIGSOFT Software Engineering Notes) in Topic 17, and WN (wireless network) in Topic 25.

In Table 2, the top ranked authors tend to have different research areas. For example, Metin Demiralp's research focus is mathematic algorithms. His work has been published in different conferences and journals specialized in microelectronics, applied mathematics, engineering, informatics and communications, and signal processing. The function of the latent variable of community is to group the authors with similar topic distributions into one community. This function can better discover authors with similar research interests and therefore can be used to make personalized recommendations. For example, according to the topic distribution of Community 2, its main topic is Topic 1 (network system), Topic 15 (parallel and



(a) Author-community distribuiotn

(b) Community-topic distribution

Fig. 5. probability distribution in Arnetminer in 2008–2010. (a) Author–community distribution and (b) community–topic distribution.

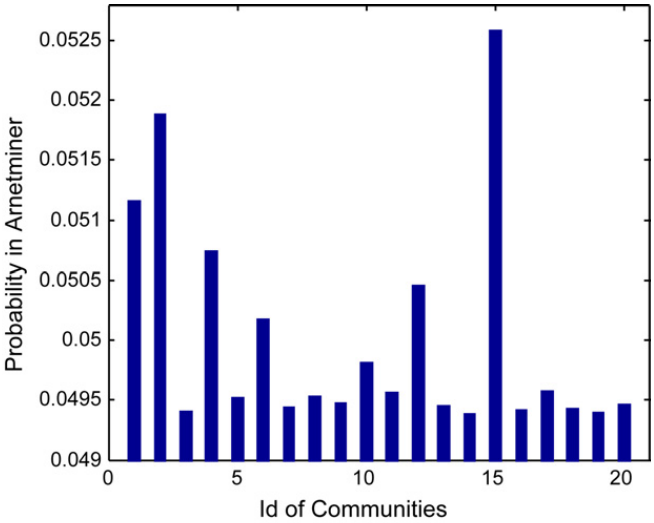


Fig. 6. Community distribution of Arnetminer in 2008–2010.

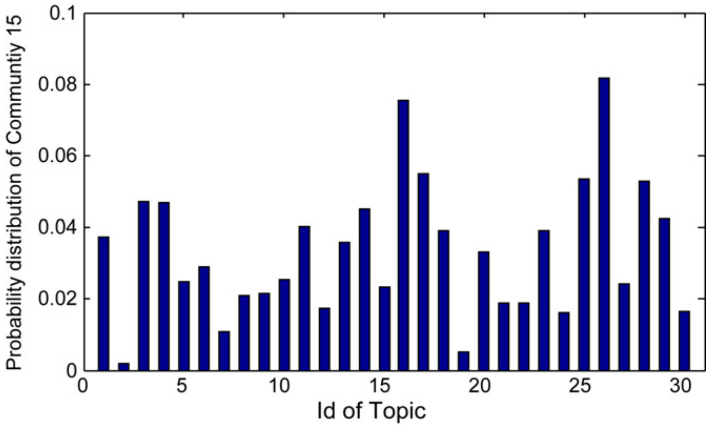
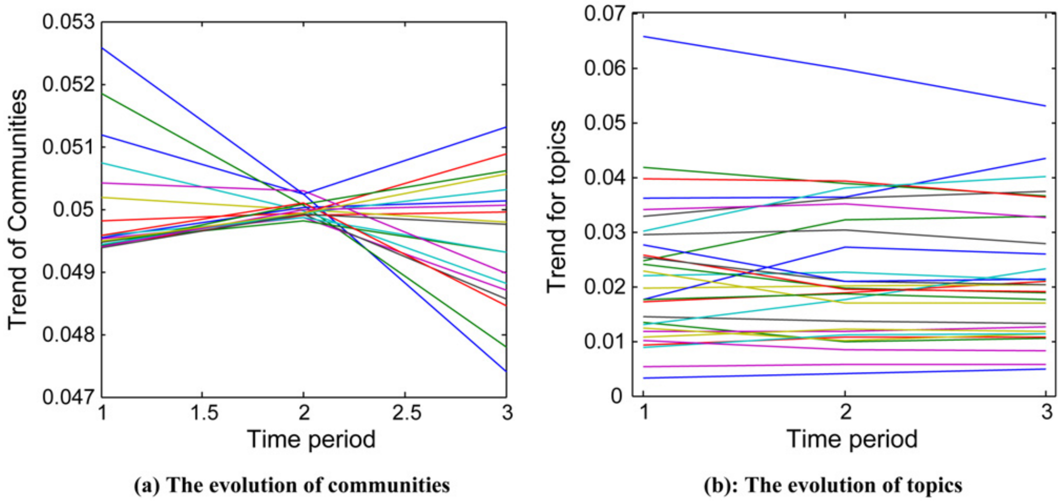


Fig. 7. Topic distribution in Community 15 in Arnetminer.

Table 2
The description of Topic 16, 17, 25, 26, 28.

Topic	16	17	25	26	28
Word	Imaging 0.049500 recognition 0.020414 detection 0.017042 video 0.014272 feature 0.013188 segmented 0.0127366	Software 0.0462338 development 0.0150655 engine 0.0144427 oriented 0.0143241 architecture 0.0121295 component 0.00987569	Networks 0.0834684 wireless 0.0353384 mobile 0.0217488 sensor 0.0214399 routing 0.0129979 protocol 0.0123287	Information 0.0240355 management 0.017254 systems 0.0149435 business 0.00921228 knowledge 0.00852213 communication 0.00834209	Computer 0.0188073 algorithm 0.0142956 polynomial 0.0141056 algebras 0.00954637 linearization 0.0093564 approximation 0.00840656
Conference	IEEE PAMI 0.0357286	IEEE Software 0.0167132	IEEE/ACM Networking 0.0308232 CN: CTN 0.0302538 WN 0.0184785 CC 0.015787	SS	JCAM 0.0406855
	PRL 0.0303648 PR 0.0229706 SP 0.0145159	ACM SSEN 0.0163849 JSS 0.0139675 IEEE SE 0.0122365		Computer 0.0168215 JASIST 0.0139223 IM 0.0131371	DM 0.0252367 JSC 0.0250928 JCTS 0.0226939 TCS 0.0226459

Metin Demiralp 0.000675616 Nikos Fakotakis 0.000411245 Yang Liu 0.00038187.
Virginie Govaere 0.000323121 Michael McAleer 0.000323121 Marc Moonen 0.000323121.
Ibrahim Busu 0.000293746 Nico Mastorakis 0.000293746 Thierry Martin 0.000293746.
Zhong Liu 0.000293746.



(a) The evolution of communities

(b): The evolution of topics

Fig. 8. The evolution of communities and topics during three time periods in Arnetminer. (a) The evolution of communities and (b) the evolution of topics.

distributed systems), Topic 25 (wireless network), Topic 16 (image recognition), Topic 8 (intelligence system and semantic), and Topic 11(fuzzy, neutral, system control). Authors who have high ranks in Community 2 may be interested in publishing articles with several of those topics. In CTM, each author has a community distribution and each community has a topic distribution. The assignment of an author to a community is based on the similarity between the topic distribution of an author and the topic distribution of a community. This is different compared to other existing LDA models. In other LDA models, each author also has a topic distribution. But authors are only grouped based on their probabilities on a single topic rather than being grouped by their probability distribution over all topics. In other words, other LDA models can automatically define a topic by using a set of words and their probabilities in that topic, while CTM can automatically define a community by using a set of topics and their probabilities in that community.

5.2. Result analysis from the dynamic perspective

DCTM has the built-in functionality to simultaneously track the temporal changes of topics and community structures, which can identify the hidden dynamic relationships between topics and communities. Here, DCTM was tested on the Arnetminer and Twitter datasets to unveil their community evolution patterns.

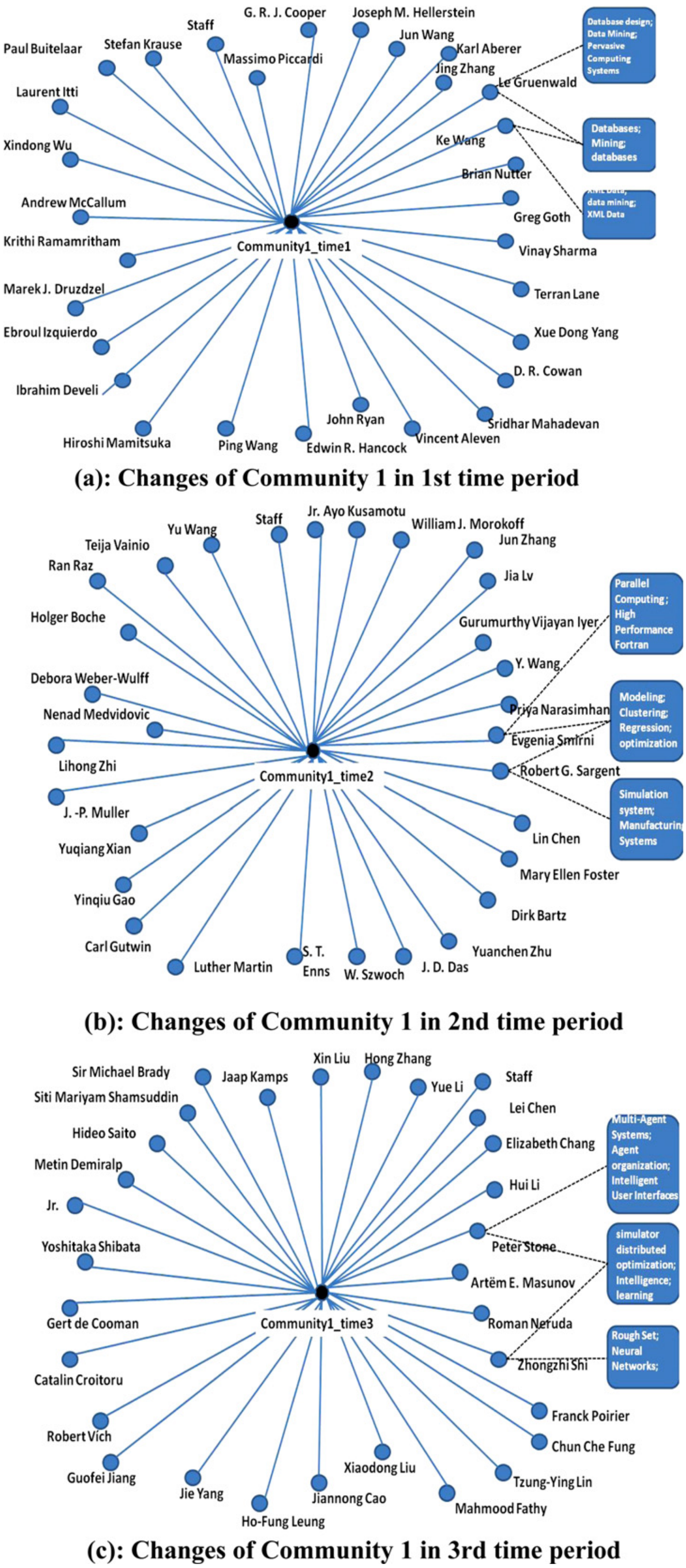
5.2.1. The features of community evolution in Arnetminer

Fig. 8 displays the evolution of all communities and topics along three time periods in Arnetminer. Fig. 8a displays the changes of all authors' preferences for each community. In period 1 and 3, authors have significant preferences for some communities, while in period 2, this phenomena is not obvious. In Fig. 8b, most of the topics exhibit a smooth increase or decrease along the time. Among them, Topic 15 (parallel computation/distribution systems) is significantly higher than other topics during all three time periods.

CTM ranks authors and topics in each community. The ranking of a topic in a community depends on the probability that the topic is assigned to the community. The ranking of an author in a community depends on two factors: first, the active level of the author during a certain time period; second, the ranking of the representative topics of the author in the community. Intuitively, if an author writes a lot of papers in the highly ranked topics in a community, the author tends to be ranked high in that community. In our experiment, we selected top 100 ranked authors in each community, and found that few top ranked authors remain the same in a community across all three different time periods. The main reason is that the ranking is based on active level of each author in a certain time period but not the influence of that author in related research area. In other words, even the ranking of a well-known author in some community will drop as long as the author does not write as many papers as before.

Another finding is that highly ranked topics for a community do not change significantly along the time. Specifically, in Community 1, the top 3 ranked topics (i.e. Topic 9: Database, 14: Machine Learning, 25: Network and Wireless) in the first time period still ranked relatively high during the next two time periods. However, the overall topic distribution is still changing and the ranking of some topic does not remain the same. For instance, in the third time period, Topic 1 (i.e. information management is ranked much higher than the previous two time periods).

Fig. 9 displays the changes of its top ranked authors and topics in Community 1. The topic distribution of Community 1 does not change significantly for the entire time periods (For example, the representative topics in blue box is mainly about database, machine learning, and clustering algorithm in first time period; distribute system, system performance in second time period; and user interface, agent, optimization, and intelligence in the third time period), while the composition of top ranked authors is experiencing relatively big changes. Very few authors remain in the same community over the three time



(a): Changes of Community 1 in 1st time period

(b): Changes of Community 1 in 2nd time period

(c): Changes of Community 1 in 3rd time period

Fig. 9. Changes of Community 1 during three time periods in Arnetminer. (a) Changes of Community 1 in 1st time period, (b) changes of Community 1 in 2nd time period and (c) changes of Community 1 in 3rd time period.

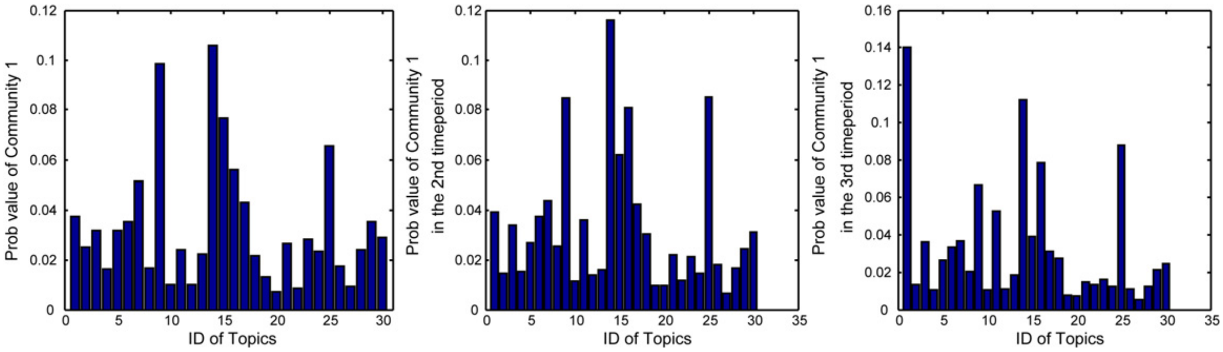


Fig. 10. Topic distribution of Community 1 over three time periods.

periods. The probability of an author for a given community is associated with the yearly productivity of this author. For example, for two authors with similar topic distributions, if an author published more papers than the other in a certain time period, he will be ranked higher.

Another finding is that highly ranked topics for a community do not change significantly along the time. As can be seen in Fig. 10, the top 3 ranked topics (i.e. Topic 9: Database, 14: Machine Learning, 25: Network and Wireless) in the first time period still ranked relatively high during the next two time periods. However, the overall topic distribution is still changing and the ranking of some topic does not remain the same. For instance, in the third time period, Topic 1 (i.e., information management) is ranked much higher than the previous two time periods).

5.2.2. *The features of community evolution in Twitter*

Unlike the academic community from the Arnetminer data, it is hard to find representative users of Twitter from each community and analyze their interests. Thus, we focus on how the overall composition of the community evolves over time and around different topics. DCTM was applied to the Twitter data to detect the evolution of a community and its relation to the topic of the community. Two representative communities from the Twitter dataset were selected for the comparative analysis.

Figs. 11 and 12 show the community evolution of the Community 0 and Community 13 during three periods of time (i.e. July, August, and September). Each star-network represents the constitution of the community at a specific time period, with the central node representing the community and time ID, and the surrounding nodes representing the users. The edge between the central node and surrounding nodes represents the member relationship. Figs. 13 and 14 highlight the key words of the corresponding community whose size is proportional to its frequency in that community. These highlighted key words can be viewed as the main topics in the corresponding community.

Many users in Community 0, like Job_Universe and photosrus, consistently belong to the community from period 1 to period 3. During each time period, some new users join the community. The consistent users constitute the core of Community 0 while the new users lie around the circumference of Community 0. However, the evolution of Community13 is quite different. From time period 1 to time period 2, some users act as a bridge to connect Community 13 during the two

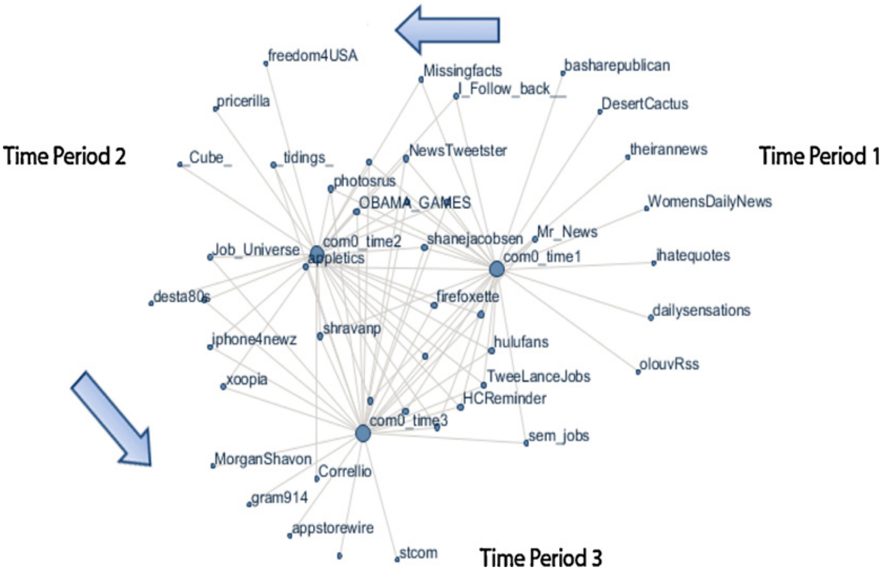


Fig. 11. Evolution of Community 0 in Twitter.

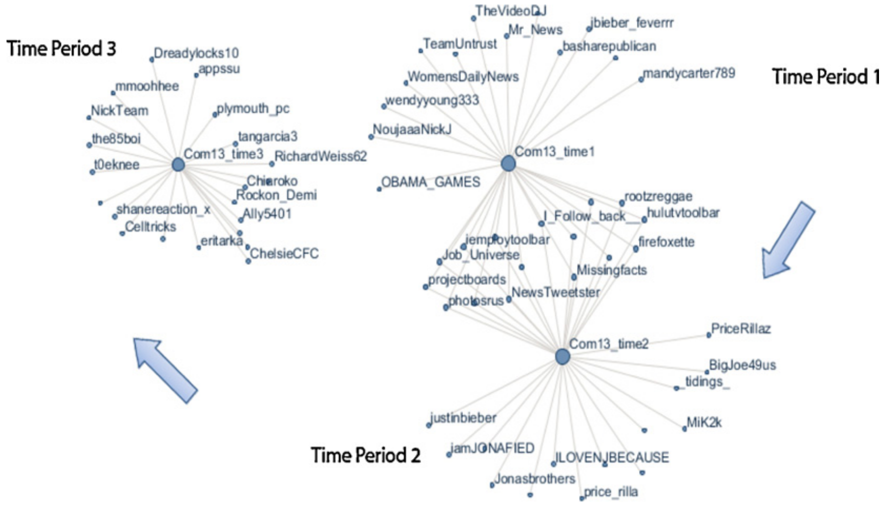


Fig. 12. Evolution of Community 13 in Twitter.

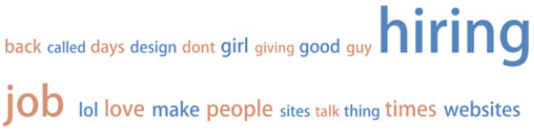

back called days design dont girl giving good guy hiring
job lol love make people sites talk thing times websites

Fig. 13. Highlighting words for Community 0 in Twitter.

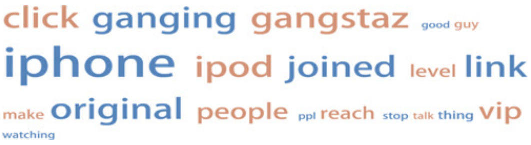

click ganging gangstaz good guy
iphone ipod joined level link
make original people ppl reach stop talk thing vip
watching

Fig. 14. Highlighting words for Community 13 in Twitter.

periods of time, which is very similar to Community 0. But during the time period 3, Community 13 is flushed with new users and lost many old users. The size of the community shrinks during this time period as well.

The topic of Community 0 is primarily about job hiring. It is a stable topic that may not exhibit traffic bursts and is regularly tweeted in Twitter. Therefore, the community organized around this type of topic has relatively fixed members. The topic of Community 13 focuses on iPhone. The launch of iPhone 4 in the end of July garnered a lot of attention in the marketplace. From the beginning of July to the middle of August, many tweets about iPhone emerged. These types of users can be called “social precursors” who are willing to initiate the use of new technology and spread it out to other members. However, when the iPhone 4 becomes more common, the tweets about this topic drop significantly and eventually diminish. Only a very small number of tweeters who are aware of this topic much later form this isolated Community 13 after the middle of August. This type of topic is “hot but transient” and is generally called a “burst”.

The dynamics of popularity of different topics can be directly detected by the DCTM. Among the 30 topics extracted by the DCTM from the Twitter data, some representative social topics were selected for discussion and their popularity tendency was plotted in Fig. 15.

The World Cup was a popular topic during its final match but lost its popularity soon after the match is over. The discussion of the iPhone became extremely intense right after its launch at the end of July and gradually cooled down when it was replaced by its next generation. The topic about the Australian election reached its peak around the end of August when the final results were publicized. After that, its popularity fell, but not very sharply, implying some continuous discussion was still going on. Similarly, the topic of Indonesian Independence Day reached its peak at almost the same time as the Australian election and remained at its peak for quite a while. The topic popularity of President Obama’s healthcare plan dropped linearly throughout the whole period. The topic of the family life, which is a consistent topic in Twitter, showed only 0.8% fluctuations in popularity during the whole period.

6. Discussion

6.1. The dynamic function of DCTM

The dynamic function of DCTM can estimate the current topic and community distributions based on the priori knowledge from the previous time period. Here the experiment was designed to demonstrate its dynamic function. The entire Twitter

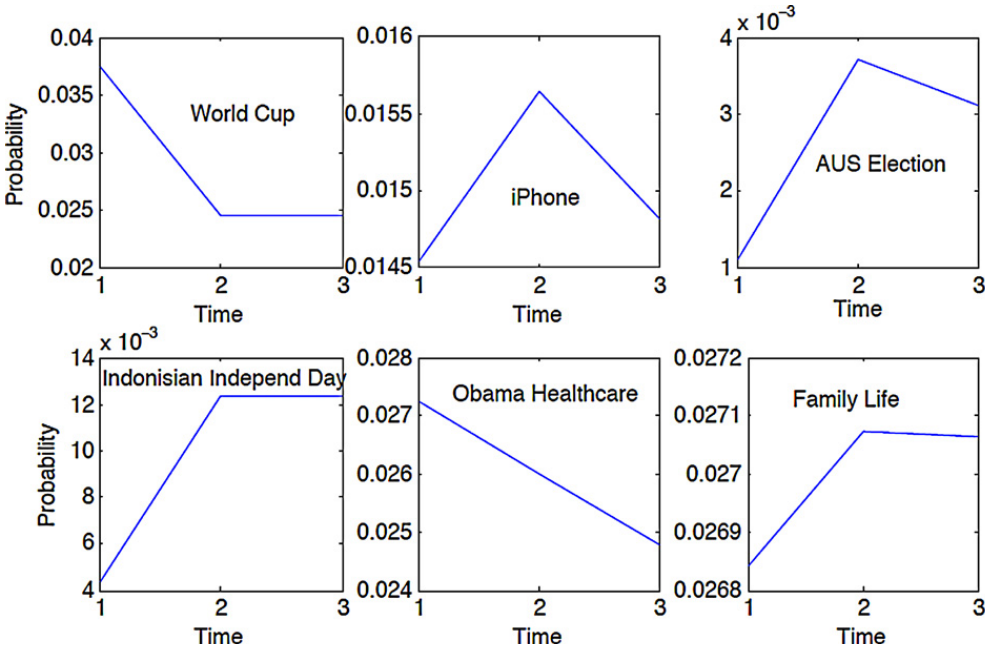


Fig. 15. Topic temporal tendency for representative words in Twitter.

Table 3
F1-measure of CTM and DCTM.

Arnetminer	Index	CTM	DCTM
Time period 2 (2004–2007)	F1-measure	0.1051	0.1102
Time period 3 (2008–2010)	F1-measure	0.0944	0.1073

dataset, and a subset of the Arnetminer that includes 10,000 publications, 5307 conferences, 166,774 authors, and 26,617 words from 2000 to 2010, were selected. The experiment contains the following steps:

Step 1: 10% of the papers from each dataset (including Arnetminer and Twitter) of each time period were randomly selected as a testing data; the rest was used as a training data.
Step 2: DCTM was applied to the entire training dataset and generated the author–community, community–topic, topic–word, and topic–conference distribution matrixes for each time period.
Step 3: The training data was divided into three time periods and the CTM was applied on the training data in each time period.
Step 4: For each time period, the results of CTM and DCTM were used to recommend a conference or journal (here a hashtag in tweets was viewed as a conference or journal) for each paper or tweet in the testing data separately. F1-measure was used to evaluate the recommendation results of the CTM and DCTM correspondingly.

Table 3 shows that the dynamic function can significantly improve the performance of the DCTM for conference and journal recommendations. For each time period, it can use priori knowledge to estimate the new dataset and adjust the results. Unlike the performance of the CTM and DCTM in the academic world, DCTM does not outperform CTM on the Twitter data, implying that historical information does not improve the task of recommending hashtags for tweets. There are two explanations: (1) a hashtag is more ephemeral than a conference. Most conferences are held yearly, while most hashtags only occurred frequently during a specific period of time; and (2) the informality and inconsistency of hashtag usage is another reason. It is common that many hashtag users ignore the original meaning of a particular hashtag. An obvious example is the overwhelming amount of spam tweets with hot hashtags. It remains as our future work on the optimization of both models for twitter recommendation.

We applied “Statistical Significance Test” to further prove the advantage of dynamic mechanism. In order to get enough samples, we set the number of time periods 10 (each year can be seen as a timeperiod) and re-did the experiment and calculated F1-measure of both CTM and DCTM in each timeperiod. *t*-test is applied to compare the results of CTM and DCTM, and the final *p*-value and deviation value can be seen as below in Table 4:

As can be seen in Table 4, the *p*-value of *t* test is smaller than 0.05, which means that there exists a statistically significant difference in performance between DCTM and CTM. The average deviation is bigger than zero, which means that DCTM outperformance CTM on F1-measure.

Table 4
p-Value for model comparison *t* test.

	p-Value	Average deviation
DCTM vs CTM	<0.05	+0.0058

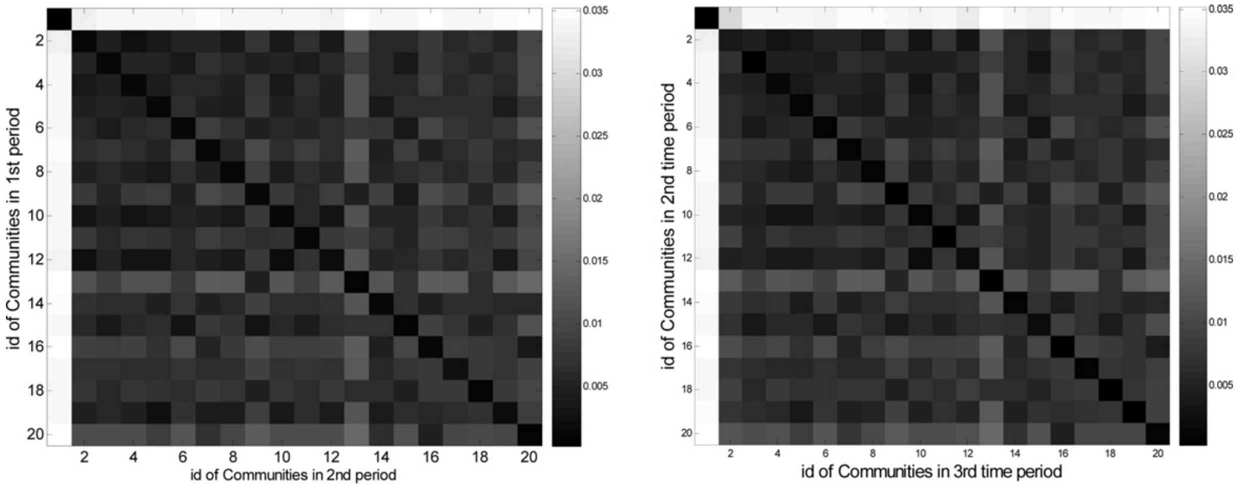


Fig. 16. Community similarity between each adjacent time periods based on DCTM.

6.2. Dynamic analysis of topic and community

Griffiths and Steyvers (2004) pointed out that an important part of realizing dynamic topic model is to build up the consistency for latent variables between adjacent timestamps. The dynamic function of DCTM can automatically generate communities and topics at the first time point and guarantee the consistency of latent variables for other time points. Based on the Arnetminer dataset, the similarity of all communities and topics between each two adjacent time periods was calculated and displayed in below heat-maps (Figs. 16 and 17):

In Figs. 16 and 17, the dark color means that two variables have a high similarity. All the heat-maps exhibit a high similarity on the diagonal, which means that the same latent variables can be assigned to a unique id through the whole time period. To compare with results generated by CTM based on the same dataset (see Fig. 18), it is clear that there is no consistency for communities for different time periods. DCTM demonstrates the clear advantage on identifying the consistency for the latest variables.

6.3. Community content and structure analysis

Most community detection algorithms are based on the graph topology of nodes and edges. The members in a community identified by CTM and DCTM demonstrate the strong topic similarities. Therefore, authors in such community may not

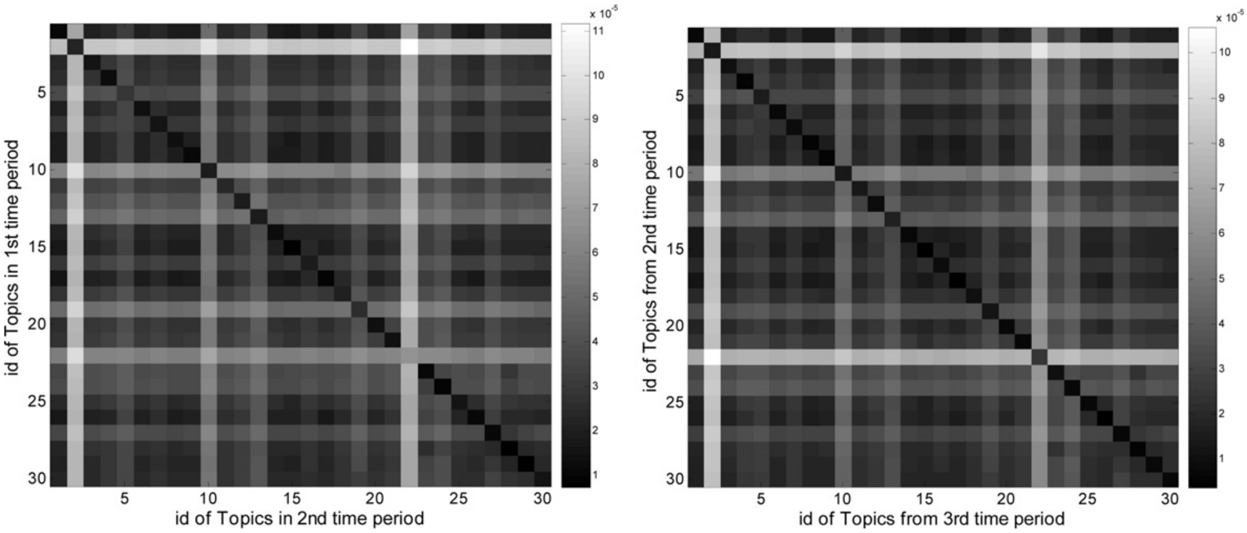


Fig. 17. Topic similarity between each adjacent time periods based on DCTM.

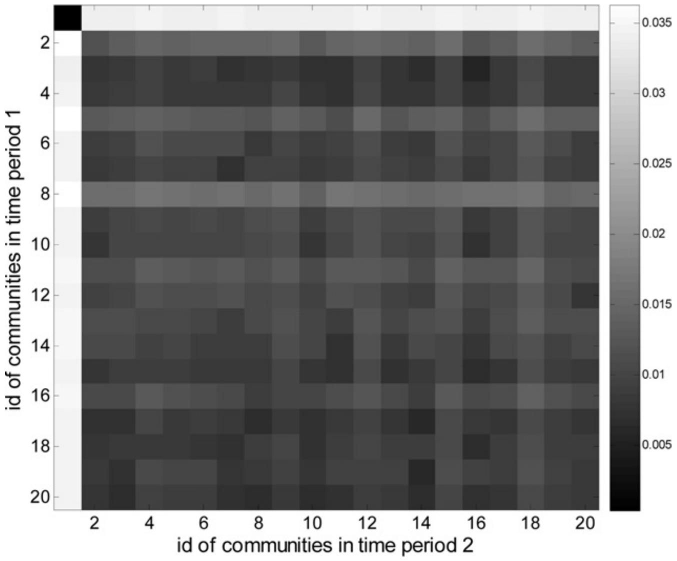


Fig. 18. Communities similarity between two adjacent time periods based on CTM.

Table 5
Conductance for five communities derived from DCTM.

Community id	14	4	20	9	11
Conductance	0.75399	0.76475	0.77233	0.79515	0.79544

Table 6
Conductance for five communities derived from the Girvan–Newman approach.

Community id	7	17	8	16	18
Conductance	0.02229	0.03265	0.03541	0.03684	0.03775

Table 7
Average *sKL* for five communities derived from DCTM.

Community id	14	4	20	9	11
Average *sKL*	0.0435	0.0379	0.0446	0.0382	0.0433

coauthor with each other but do share common topic interests. Conductance was used to measure the quality of different communities, which is defined as (Leskovec et al., 2008):

$$f(C) = \frac{s_c}{2m_c + s_c} \qquad (4)$$

C denotes the set of nodes in a community, m_c as the number of edges in C, and $s_c = |\{(u, v)|u \in C \& v \notin C\}|$ is the number of all (u, v) that satisfies the condition. According to the definition of conductance, a community of high quality should have a low conductance value. Girvan–Newman community detection algorithm was applied to the coauthor network from the small dataset of Arnetminer (including 10,000 papers) (Girvan & Newman, 2002). The detected communities were compared with the communities derived from DCTM. Tables 5 and 6 list the conductance of top five communities derived from the Girvan–Newman approach and DCTM correspondingly.

The conductance of communities identified by DCTM is higher than communities derived by the Girvan–Newman approach, which means that the number of coauthor connections in a DCTM community is fewer than that in a Girvan–Newman community. Average *sKL* divergence was calculated for the 500 authors in DCTM communities and 500 authors in Girvan–Newman communities (Rosen-zvi et al., 2008). The results are summarized in Tables 7 and 8.

The low *sKL* divergence means the nodes in a community have high topic similarity. The average *sKL* in Table 7 is higher than that in Table 8 indicates that authors in Girvan–Newman community tent to share similar topics than those in the DCTM

Table 8
Average *sKL* for five communities derived from the Girvan–Newman approach.

Community id	7	17	8	16	18
Average *sKL*	0.0252	0.0326	0.0192	0.0272	0.0308

community, because the co-author relationship reflects a strong semantic connection among different nodes. However, the results in Table 7 still indicate that DCTM can discover authors with relative high similarity of topic distribution in a detected community, while those authors may have few co-author relationships compared with the communities derived from Newman–Girvan (in Tables 5 and 6, the authors in the DCTM community tent not to collaborate with each other). So the nodes in the community identified by the DCTM model embedded the feature of sharing similar topics but collaborating in a limited manner. That can provide meaningful recommendations for authors who would like to find potential cooperators that they do not know before.

We also calculated the average conductance of all communities detected by CTM and Girvan–Newman algorithm separately, and obtained the result of 0.03343 from 132 communities detected by Girvan–Newman algorithm, as well as 0.7742 from 20 communities detected by CTM. We found that the average conductance of communities identified by DCTM is higher than communities derived by the Girvan–Newman approach, implying that the number of coauthor connections in a DCTM community is fewer than that in a Girvan–Newman community.

7. Conclusion and future work

In this paper, we present the CTM and DCTM to detect communities and topics. The CTM contains four observed variables and two latent variables. It can discover topic features from the four observed variables and uses the relationships to define communities. The dynamic function of DCTM takes into account the temporal continuity between consecutive timestamps that ensures the consistency for each community and topic during the whole time period. Experiments show that the CTM can find communities sharing similar topics, while the DCTM can identify the dynamic features of communities and topics. In the future, we will integrate a supervised model into the DCTM in order to capture the highly cited authors instead of highly productive authors. In order to improve computational efficiency, we will consider adopting parallel computing technology to accelerate the process.

Acknowledgements

This work is supported by NIH-funded VIVO project (NIH grant U24RR029822).

Daifeng Li is funded by China National Natural Science Foundation (70971083), the Graduate Innovation Fund of Shanghai University of Finance and Economics (cxjj-2008-330), the 2009 Doctoral Education Fund of Ministry of Education in China (20090078110001), the NIH VIVO project (uf09179) and Chinese National Key Foundation Research (60933013).

Jie Tang is supported by the Natural Science Foundation of China (No. 60703059), Chinese National Key Foundation Research (No. 60933013), and National High-tech R&D Program (No. 2009AA01Z138).

References

Andersen, R., Chung, F., & Lang, K. (2008). Local partitioning for directed graphs using pagerank. *Internet Mathematics*, *51*, 3–22.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *The Proceedings of the 23rd international conference on machine learning (ICML2006)* Pittsburgh, PA, USA, (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chang, J., & Blei, D. M. (2009). Relational topic models for document networks. In *Proceedings of the 12th international conference on artifical intelligence and statistics (AISTATS)*

Clauset, A., Moore, C., & Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*, 98–101.

Ding, Y. (2011). Community Detection: Topological vs. Topical. *Journal of Informetrics*, *5*, 498–514.

Erdos, P., & Renyi, A. (1959). On random graphs. *Publications Mathematicae*, *6*, 290–291.

Flake, G., Tarjan, R., & Tsioutsiouliklis, K. (2003). Graph clustering and minimum cut trees. *Internet Mathematics*, *14*, 385–408.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *9912*, 7821–7826.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *6*(101 Suppl. 1), 5228–5235.

He, Q., Chen, B., Jian, P., Qiu, B., Mitra, P., & Giles, C. L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *CIKM 09* HongKong, China, November 2–6.

Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2010). Online multiscale dynamic topic models. In *The 16th ACM SIGKDD conference on knowledge discovery and data mining*.

Leskovec, J., Lang, K., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the nineteenth international world wide web conference* North Caroline, USA.

Leskovec, J., Lang, K., Dasgupta, A., & Mahoney, M. (2008). Statistical properties of community structure in large social and information networks. In *WWW'08: Proceedings of the 17th international conference on world wide web* (pp. 695–704).

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., et al. (2010). Community-based topic modeling for social tagging. In *The 19th ACM international conference on information and knowledge management (CIKM2010)* Toronto, Canada, October 26–30.

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community. In *Paper presented at proceedings of the 26th annual international conference on machine learning*.

Lu, C., Hu, X., Chen, X., & Park, J. (2010). The topic-perspective model for social tagging systems. In *The 16th ACM SIGKDD conference on knowledge discovery and data mining 7*.

Milgram, S. (1967). The small world problem. *Psychology Today*, *2*, 60–67.

Nallapati, R., & Cohen, W. (2008). Link-PLSA-LDA: A new unsupervised model for topics and influence in blogs. In *Proceedings of international conference on weblogs and social media ICWSM'08* (pp. 84–92).

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *982*, 404–409.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *1019*, 2658–2663.

Rosen-zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2008). The author–topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence* (pp. 487–494). Virginia: AUAI Press.

Si, X., & Sun, M. (2009). Tag-LDA for scalable real-time tag recommendation. *Journal of Computational Information Systems*, *1*, 23–30.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the fourteenth ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'2008), 99* (pp. 0–998).

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of 2008 IEEE international conference on data mining ICDM'2008* (pp. 1055–1060). Washington, DC: IEEE Computer Society.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence [UAI]*.

Zhou, D., Manavoglu, E., Li, J., Giles, C. L., & Zhai, H. (2005). Probabilistic models for discovering e-communities. In *WWW 06* Edinburgh, Scotland, May 23–26.

Zhu, S., Yu, K., Chi, Y., & Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *SIGIR, 07* Amsterdam, The Netherlands, July 23–27.