IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 14, 2019, accepted June 5, 2019, date of publication June 10, 2019, date of current version June 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921824

Dynamic Topical Community Detection in Social Network: A Generative Model Approach

YUNLEI ZHANG, BIN WU, NIANWEN NING, CHENGUANG SONG, JINNA LV
Beijing Key Laboratory of Intelligence Telecommunications Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Bin Wu (wubin@bupt.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0831500.

ABSTRACT Social networks that are dynamic contain rich network structure and content information. In dynamic networks, it is necessary to discover communities and their topical meanings. However, existing methods either only discover communities with ignoring their topical meaning in dynamic networks, or they discover communities and their topics in static networks. In this paper, we identify the problem of dynamic topical community detection and propose a dynamic topical community detection (DTCD) method to detect communities and their topical meanings in dynamic networks. The DTCD is a generative model integrating network structure, text, and time. The DTCD considers a community as a mixture of topics and generates the neighbors and documents of the node and their time stamps at the same time via the community. The latent variables are learned by collapsed Gibbs sampling. The DTCD not only can find communities and their topics, but also capture the temporal variations of communities and topics. The experimental results on two real-world datasets demonstrate the effectiveness of DTCD.

INDEX TERMS Social network, dynamic community detection, user generated content, generative model, collapsed Gibbs sampling.

I. INTRODUCTION

Networks are ubiquitous in real-world. In recent years, social network has become a fundamental tool for communication and obtaining information, such as twitter, weibo and wechat. Community structure is an important property of these social networks. Users may conduct more similar behavior in the same community than ones do in the different communities. Obviously, these social networks are dynamic. Therefore, the communities are also changing when the social network is evolving. Discovering the dynamic community structure [1] is an important task in social network analysis. It is helpful to other data mining tasks, such as information diffusion [2], influence maximization [3], group recommendation [4], etc. Recently, many researchers have proposed many methods for dynamic community detection, e.g., [5]–[17]. However, these methods only detect dynamic community based on network structure. Surely, these methods find communities in which the members have dense connections between them, and have sparse connections with the members of other different communities. In other words, these communities are defined at structural level. However, in real-world situation, besides there are links among users in social network, the users also publish texts, such as publishing papers in academic cooperation network, posting texts in weibo social network. Social network containing time and content information is termed as dynamic information network.

We list two exemplar social networks as follows:

- **Online forum social network**: The users in the forum are taken as nodes, and the interactions between them are taken as edges in the network. A user publishes a post to launch a discussion which may be commented by other users. The contents published or commented by the users are taken as node content in the network.
- **Academic cooperation social network**: The authors are taken as nodes, and the cooperations among authors are taken as edges in the network. The node content consists of the content of the papers published by the corresponding authors.

These existing approaches only based on network structure may work well in networks with communities exhibiting internal dense connections. However, the users with similar contents may represent that they have the similar interests, whereas there are sparse connections between them.

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

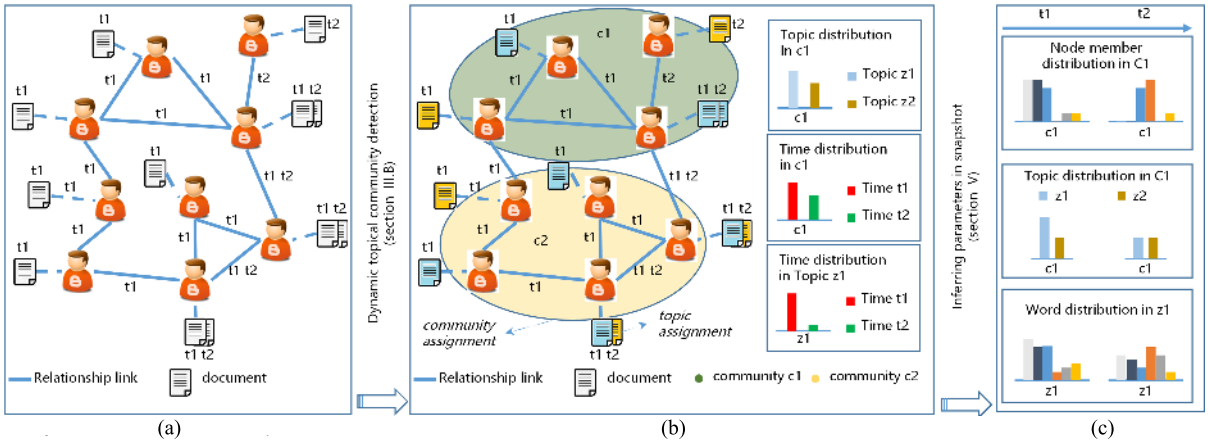


FIGURE 1. **Overview of dynamic topical community detection. (a) Dynamic network with documents input. (b) Dynamic topical community detection. (c) Distributions.**

Moreover, there are latent topics (refer to definition 3) which is considered as multinomial distribution over words in these contents like in [18]. These methods only based on links cannot gather these users as a community properly. Although the users with the same interests in some topics have sparse connections, they should be considered as a community in social networks with node content. The methods adopt only network structure may fail to detect community at topical level. To discover both structural and topical level communities, both network structure and node content should be taken into consideration together. In this paper, we propose a dynamic topical community method which integrates network structure and node content.

Figure 1 shows an overview of dynamic topical community detection process. The input is dynamic network with node content. And it aims to reveal the latent communities, topics and their temporal variations.

Dynamic topical community detection is challenging. Both communities and topics are hidden and changing over times. Methods to detect communities and topics separately cannot reveal correlation between them. Although in recent years some works [19], [20] have been proposed by integrating network structure and node content, they do not take the correlation between them into consideration. In this paper, we not only detect communities and extract topics by considering the correlation between them, but also characterize their temporal variations.

To detect communities and extract topics simultaneously, a generative model is proposed, called DTCD(Dynamic Topical Community Detection), and it integrates the network structure, node content and time stamp in a unified way. The community and topic are modeled as latent variables, and a generative process is proposed to observe network structure, text, and time stamp to find the communities, to extract topics and to characterize their temporal variations. And we propose a collapsed Gibbs sampling inference method to obtain the value of the latent variables. Finally, the community membership and topic membership in snapshot network are obtained by inferring parameters based on the value of the latent variables.

Based on the DTCD model, we can detect communities and their temporal variations, extract topics and their temporal variations, infer the relationship between community and topic, infer communities and topics in each snapshot network.

The basic idea of our proposed model consists of some assumptions as follows.

- The users may affiliate to different communities, and they are modeled as multinomial distribution over community, and the entries in the distribution represent the degree that the user participates in the communities.
- Previous methods assume that a community only is interested in one topic which is not proper. In this paper, we assume that one community may be interested in more than one topic. Therefore, each community not only is considered as a multinomial distribution over users indicating the users' significance in the community, but also is considered as a multinomial distribution over topics indicting the interesting topics of the members in the community.

To summarize, we make the following contributions:

- The problem of dynamic topical community detection is defined. It detects communities and topics in dynamic network, and paves the road to explore the relationship between community evolving and topic changing.
- The hidden communities and topics are uncovered in our proposed model, and the relationship between them are revealed. The temporal variations of them are also captured.
- Our model integrates the network structure, text and time stamp in a unified way. It provides an idea for using more information to detect community and topic accurately. Our method iteratively detects community and topic which improve each other.

The rest of this paper is organized as follows: Section II reviews related literatures. Section III formulates the problem and introduces the model. Section IV describes the inference method. Section V describes the inference of parameters in snapshot network. Section VI evaluates the solution. Finally, we conclude this work in section VII.

II. RELATED WORK

In this section, we will review the related works with dynamic topical community detection. It consists of two parts: dynamic community detection and community detection with node content. The former part mainly focuses on dynamic community detection by using network structure which ignoring node content containing community's semantic information. The later part mainly focuses on community detection in static network by using network structure and node content.

A. DYNAMIC COMMUNITY DETECTION

An agglomerative hierarchical clustering was proposed by Hopcroft *et al.* [5] to detect evolving communities. Backstrom *et al.* [6] conducted experiments to answer the questions of community membership, growth and evolution. Sun *et al.* [7] proposed a method discovering evolving communities by means of information compression. Palla *et al.* [8] proposed a method discovering evolving communities by means of clique percolation. Tang *et al.* [9] analyzed a multi-mode network via temporal information and detected community evolution. Asur *et al.* [11] proposed an event-based method to characterize dynamic relationship between nodes and communities. Alvari *et al.* [21] proposed a game-theoretic approach for community detection in dynamic social network. Wang *et al.* [12] proposed a unified random walk method to detect communities in dynamic networks by integrating network structure, node content and edge content. HOCTracker [13] is a unified framework which finds out the evolution patterns of hierarchical and overlapping communities in online social networks. Hu *et al.* [14] proposed a method to track dynamic communities and their evolutionary behaviors by exploring the local views of nodes that change. Ma *et al.* [15] proposed a semi-supervised evolutionary nonnegativematrix factorization (sE-NMF) for detecting dynamic communities by incorporating a priori information into ENMF. Zhou *et al.* [16] proposed a multiobjective discrete cuckoo search method to detect dynamic community by optimizing modularity and NMI. Cheng *et al.* [17] proposed a novel method for detecting new overlapping community in evolving networks by node vitality for modeling network evolution constrained by multiscaling and preferential attachment. Most of these above methods detect dynamic community by only using network structures. Only Wang *et al.* [12] proposed a method considering network structures and node content, but it ignores the topics in the communities.

B. COMMUNITY DETECTION WITH NODE CONTENT

In static network scenario, some researchers have improved the accuracy of detecting community by integrating node content and network structure [2], [19], [22]–[29]. Yang *et al.* [22] proposed a discriminative model to detect community by integrating network structure and node content. Qi *et al.* [24] took edge content into consideration to improve the effectiveness of the community detection in social media network. Liu *et al.* [19] jointly modeled topics and author community in Topic-Link LDA model, which combines the edge content and network structure. Zhou *et al.* [23] proposed a heterogeneous random walk method on graph augmented by node attributes to detect community more accurately. Sachan *et al.* [25] proposed generative models to jointly model the discussed topics, network structure and interaction type to find communities with topical meaning. Yang *et al.* [30] developed communities from network structure and node attributes. Ruan *et al.* [31] proposed a biased edge sampling procedure by integrating content and links, which retains edges that are locally relevant for each graph node. Then standard community discovery algorithms were used to cluster the resulting backbone graph. Hu *et al.* [26] proposed a method to discover social circles in ego network [32] by integrating node profile. Hu *et al.* [2] proposed a method to model users' information diffusion behaviors at community-level. Liu *et al.* [27] treated a network as a dynamic system and considered its community structure as a consequence of interactions among nodes by introducing the principle of content propagation and integrating the aspects of structure and content in a network naturally. To deal with network topology and semantic information on nodes simultaneously, Wang *et al.* [28] proposed a novel nonnegative matrix factorization model with the community membership matrix and community attribute matrix. He *et al.* [29] introduced a novel generative model with two closely interdependent components, one for network structure and the other for semantics via combining network structure and node content. Cai *et al.* [33] proposed a method to profile community, which includes profile in user generated content and diffusion information. Zhao *et al.* [34] proposed an overlapping community detection method, namely, latent Dirichlet allocation-based link partition (LBLP), which uses a graphical model and considers network structure and content information. Li *et al.* [35] constructed a newly higher-order attribute homogenous motif network by integrating network motif and node attributes for community detection. Nan *et al.* [36] proposed a semi-supervised non-negative matrix factorization model to discover communities by integrating network structure, node content and individual label.

All the above methods detect community in static network by using network structures and node content, but it is not trivial to extend these methods to dynamic networks.

TABLE 1. Notations.

Notations	Description
C, Z	The set of communities and topics
D, W	The set of documents and words
U, E, T	The set of users, edges and time slices
$\|A\|$	The number of elements in set A
N_u	The set of neighbour nodes of the user u
D_u	The set of documents associated with user u
W_{ud}	The set of words in d-th document associated with user u
η_c	Multinomial distribution over users specific to community c
x_{ui}	The community label of the i-th neighbour of user u
f_{ui}	the i-th neighbour of user u
t_{ui}	The time slice label of the i-th neighbour of user u
y_{ud}	The community label of the d-th document associated with user u
z_{ud}	The topic label of the d-th document associated with user u
t'_{ud}	The time slice label of the d-th document associated with user u
w_{udj}	The j-th word in the d-th document associated with user u
Ω_c	Multinomial distribution over time slices specific to community c
Ψ_z	Multinomial distribution over time slices specific to topic z
ψ_c	Multinomial distribution over topics specific to community c
ϕ_z	Multinomial distribution over words specific to topic z
$\alpha, \beta, \gamma, \delta$	Dirichlet priors

III. DYNAMIC TOPICAL COMMUNITY DETECTION METHOD

The central task is to detect communities and topics from networks which contains network structure, user content and time slice information, and utilize them to infer communities and topics in the snapshot network.

In this section, we formulate the dynamic topical community detection problem. We then propose DTCD which is a comprehensive latent variable model to address the problem.

A. PROBLEM FORMULATION

The notations used in this paper are listed in Tab. 1.

Definition 1 (Social Network): A social network is $G = (U, E, D, T)$, where U is a set of users, E is a set of edges representing the relationship between users such as friendship, cooperation, D is a set of documents which are associated with users in U and T is a set of time slices representing the generation time of E and D.

Let D_u denote the set of documents associated with user $u \in U$; each document $d_{ij} \in D_u$ consists of a sequential list of words from a given word set; each document d_{ij} has its time stamp t_{ij}. An edge $(u, v, t) \in E$ represents there exists communication between user u and v at time t. For a DBLP network, D_u is the set of papers published by author u; $(u, v, t) \in E$ represents that author u co-authored with author v at time t.

Definition 2 (Snapshot Network): A snapshot network is a particular kind of social network G, denoted as $S = (U, E, D, T)$, where $|T| = 1$. It means that all the edges and documents in the network appears at the same time.

To take node content into consideration for topic modeling, we give the definition of topic.

Definition 3 (Topic): A topic $z \in Z$ is a $|W|$-dimensional multinomial distribution over words, denoted as ϕ_z, where each entry $\phi_{z,w}$ denotes the probability of a word $w \in W$ generated by topic z.

A topic has different popularity at different time. We characterize the changing popularity by a temporal distribution, which is defined as follows:

Definition 4 (Topic Temporal Variation): A topic $z \in Z$ temporal variation is a $|T|$-dimensional multinomial distribution over the time stamps, denoted as Ψ_z, where each dimension $\Psi_{z,t}$ is the probability of a topic z occurring at time stamp t.

In this paper, community is characterized in three aspects, including network structure, topic extracted from text(document) and time stamp. First, community consists of a set of users having more links with users in the same community and less links with users in the different community in terms of network structure. Second, community is correlated with a mixture of topics extracted from node content, and it represents the interests of the corresponding community in terms of topics. Whereas existing methods assumes one community only is interesting to one topic, which is not suitable. Finally, community is characterized by a time stamp distribution, which represents the popularity of the community in terms of time stamp. We give the definitions from these three aspects as follows:

Definition 5 (Community): A community $c \in C$ is a $|U|$-dimensional multinomial distribution over users, denoted as η_c, where each entry $\eta_{c,u}$ represents the significance of the user u in the community c.

Definition 6 (Community's Topic Profile): A community's topic profile is a $|Z|$-dimensional multinomial distribution over topics, denoted as ψ_c, where each entry $\psi_{c,z}$ represents how much interest community c has in topic z;

Definition 7 (Community Temporal Variation): A community c's temporal variation is a $|T|$-dimensional multinomial distribution over time slices, denoted as Ω_c, where each component $\Omega_{c,t}$ represents the popularity of community c at time stamp t.

In social networks, users usually are affiliated to different communities with different degree [37]. We model users in social networks in *mixed-membership* manner.

Definition 8 (User's Community Membership): A user u's community membership is a $|C|$-dimensional multinomial distribution over communities, denoted as π_u, where each component $\pi_{u,c}$ indicates user u's affiliation degree to community c.

Take community $c1$ and topic $z1$ in Fig. 1 as an example. As $c1$'s users publish more documents on topic $z1$ than topic $z2$, the resulting $\psi_{c1,z1}$ is bigger than $\psi_{c1,z2}$; as $c1$'s users interact with other users at time stamp $t1$ more frequent than at time stamp $t2$, the resulting $\Omega_{c1,t1}$ is bigger than $\Omega_{c1,t2}$. Besides, as users in network publish documents on topic $z1$ at time stamp $t1$ more frequent than at time stamp $t2$, the resulting $\Psi_{z1,t1}$ is bigger than $\Psi_{z1,t2}$. As motivated in section I, we formalize a dynamic topical community detection problem to solve in this paper.

Problem 1 (Dynamic Topical Community Detection): Given a social network $G = (U, E, D, T)$, the task of dynamic topical community detection is to infer: 1) each topic z's distribution over words ϕ_z, $\forall z \in Z$; 2) each topic z's temporal variation Ψ_z, $\forall z \in Z$; 3) each community c's distribution over users η_c, $\forall c \in C$; 4) each community c's temporal variation Ω_c, $\forall c \in C$; 5) each user u's community membership π_u, $\forall u \in U$; 6) each community c's topic profile ψ_c, $\forall c \in C$.

B. MODEL STRUCTURE

DTCD integrates over text, time and network structure in a generative model approach. It models community, topic and relationship between community and topic and the temporal variation of community and topic as latent variables. In other words, the neighbors of the user who belongs to community c are generated by community η_c, the time stamp of the user who belongs to community c is generated by community c's temporal variation Ω_c, and the words of the document which is associated with topic z are generated by the topic ϕ_z, and the time stamp of the document which is associated with topic z is generated by topic z's temporal variation Ψ_z.

Like DTCD, Community Profiling and Detection (CPD) [33] is a generative model which models network structure and node content with ignoring time stamp information; Topics over Time (TOT) [38] is a topic model which integrates text and time. DTCD inspired by both CPD and TOT is a comprehensive model which integrates network structure, text and time stamp. Moreover, DTCD better fits social network with node content and dynamic structure.

DTCD models two types of user behaviors over time, i.e., one behavior is when a user publishes a document on what topic; another one is when a user interact with whom. We assume that the user behaviors which can be observed are controlled by the latent community and topic and their temporal variations in the social network. Specifically, the topics of the documents published by users are governed by the interested topics of the community c in which the users participate (i.e., ψ_c), the words of the documents are assumed to be generated by the relationship between topic and word (i.e., ϕ_z) and when the users publish documents on topic z is controlled by the relationship between topic and time stamp (i.e., Ψ_z). Moreover, the users interact with who is governed by the relationship between community and user (i.e., η_c), and when the users interact with other users is determined by the relationship between community and time stamp (i.e., Ω_c).

Figure 2 shows the graphical structure of DTCD. DTCD models the two types of user behaviors with corresponding generative process which are separated.

C. INDIVIDUAL COMPONENTS

Figure 2 shows that DTCD consists of three blocks: the *text-time block* reveals the topics underlying node content, and characterizes their temporal variations. The *network-time block* reveals communities underlying network structure, and

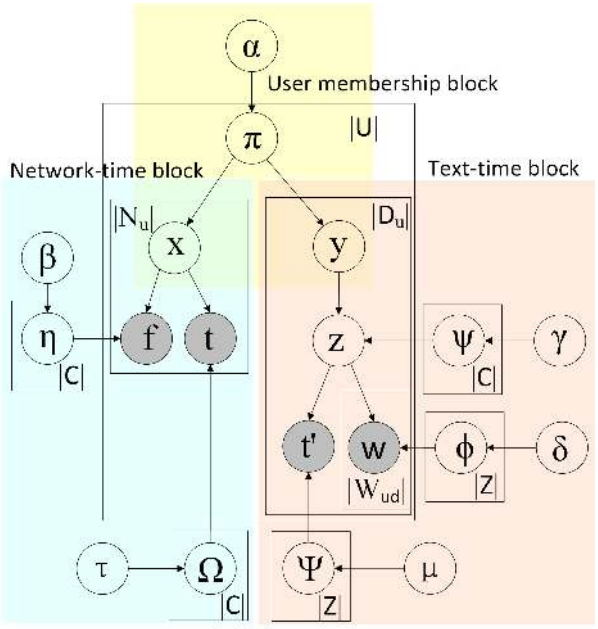


FIGURE 2. Graphical model of DTCD.

characterizes their temporal variations. The *user membership block* reveals the relationship between the users and the communities, which also provides community label to the neighbors and documents of the corresponding user.

1) USER MEMBERSHIP BLOCK

Users in social network usually affiliate to multiple communities with varying degree. We consider each user u as a community probability vector π_u. Each document $d \in D_u$ associated with user u is assigned to a community y sampled via π_u, denoting the user u is participating community y when she publishes document d. In addition, each neighbor f of user u is associated to community x, denoting node f's community membership when user u builds relationship with user f.

2) TEXT-TIME BLOCK

Each document $d \in D_u$ consists of a collection of words $\{w_{ud1}, \ldots, w_{ud|W_{ud}|}\}$, where $|W_{ud}|$ denotes the length of document d. In traditional topic models, such as latent Dirichlet allocation (LDA) [18], a document is assigned with multiple topics and each word in document has its own topic label. It is necessary to associate multiple topics to a document which consists of enough many words, such as academic papers. However, the document is short on social media network which may only undertake one topic [39]. In this paper, we assign one latent topic z which is sampled via ψ_y to a document d The words in the document d are then sampled from multinomial distribution over word specific to topic z, i.e., ϕ_z.

We use a multinomial distribution Ψ_z specific to topic z to model its temporal information which is extracted from time stamps of documents. Using multinomial distribution

can characterize the multimodal variation of the temporal variation of the topic, which can capture the character of rising and falling for many times.

3) NETWORK-TIME BLOCK

We model network structure like topic modeling in LDA. A user is considered as a document, and the collection of the neighbors as a collection of words in the document. Each $u \in U$ contains a list of neighbor $\{f_{u1}, \ldots, f_{u|N_u|}\}$, where $|N_u|$ denotes the number of neighbors of user u. The neighbors are sampled from the user distribution η_x. Like in *text-time block*, we use multinomial distribution Ω_c over time stamps to model the temporal variation of community c. The time stamp t of a neighbor f is drawn from Ω_c.

D. GENERATIVE PROCESS

The DTCD model is a generative model integrating network structure, documents associated with users and their time stamps. We summarize the DTCD model's generative process below.

(1) for each community $c \in C$:
   a) draw its $|U|$-dimensional user distribution from a Dirichlet prior parameterized by β: $\eta_c|\beta \sim Dir_{|U|}(\beta)$
   b) draw its $|Z|$-dimensional topic distribution from a Dirichlet prior parameterized by γ: $\psi_c|\gamma \sim Dir_{|Z|}(\gamma)$
   c) draw its $|T|$-dimensional time stamp distribution from a Dirichlet prior parameterized by τ: $\Omega_c|\tau \sim Dir_{|T|}(\tau)$

(2) for each topic $z \in Z$:
   a) draw its $|W|$-dimensional word distribution from Dirichlet prior parameterized by δ:$\phi_z|\delta \sim Dir_{|W|}(\delta)$
   b) draw its $|T|$-dimensional time stamp distribution from Dirichlet prior parameterized by μ:$\psi_z|\mu \sim Dir_{|T|}(\mu)$

(3) for each user $u \in U$:
   a) draw its $|C|$-dimensional community distribution from a Dirichlet prior parameterized by α: $\pi_u|\alpha \sim Dir_{|C|}(\alpha)$
   b) for each neighbor node $i \in N_u$ of user u:
      (i) draw a community assignment $x_{ui}|\pi_u \sim Multi(\pi_u)$, by user u's multinomial community distribution π_u
      (ii) draw the neighbor $f_{ui}|\eta_{x_{ui}} \sim Multi(\eta_{x_{ui}})$, by community x_{ui}'s multinomial user distribution $\eta_{x_{ui}}$
      (iii) draw time stamp of the neighbor i, $t_{ui}|\Omega_{x_{ui}} \sim Multi(\Omega_{x_{ui}})$, by community x_{ui}'s multinomial time slice distribution $\Omega_{x_{ui}}$
   c) for each document $d \in D_u$ associated with user u:
      a) draw a community assignment $y_{ud}|\pi_u \sim Multi(\pi_u)$, by user u's multinomial community distribution π_u
      b) draw a topic assignment $z_{ud}|\psi_{y_{ud}} \sim Multi(\psi_{y_{ud}})$, by community y_{ud}'s multinomial topic distribution $\psi_{y_{ud}}$
      c) draw each word $w_{udj}|\phi_{z_{ud}} \sim Multi(\phi_{z_{ud}}), \forall j = 1, \ldots, |W_{ud}|$, by z_{ud}'s multinomial word distribution $\phi_{z_{ud}}$
      d) draw time stamp of the document d associated with user u, $t'_{ud}|\Psi_{z_{ud}} \sim Multi(\Psi_{z_{ud}})$, by topic z_{ud}'s multinomial time slice distribution $\Psi_{z_{ud}}$

As shown in above process, the posterior distributions of communities and topics depend on the information from three modalities, network structure, text and time. DTCD parameterization is as follows:

$$\begin{aligned}
\eta_c|\beta &\sim Dir_{|U|}(\beta) \quad \psi_c|\gamma \sim Dir_{|Z|}(\gamma)\\
\phi_z|\delta &\sim Dir_{|W|}(\delta) \quad \pi_u|\alpha \sim Dir_{|C|}(\alpha)\\
x_{ui}|\pi_u &\sim Multi(\pi_u) \quad f_{ui}|\eta_{x_{ui}} \sim Multi(\eta_{x_{ui}})\\
t_{ui}|\Omega_{x_{ui}} &\sim Multi(\Omega_{x_{ui}}) \quad y_{ud}|\pi_u \sim Multi(\pi_u)\\
z_{ud}|\psi_{y_{ud}} &\sim Multi(\psi_{y_{ud}}) \quad w_{udj}|\phi_{z_{ud}} \sim Multi(\phi_{z_{ud}})\\
t'_{ud}|\Psi_{z_{ud}} &\sim Multi(\Psi_{z_{ud}})
\end{aligned}$$

IV. MODEL INFERENCE

A. COLLAPSED GIBBS SAMPLING

It is hard to inference parameters exactly in DTCD. In this paper, Gibbs sampling is used to infer parameters in DTCD approximately. In DTCD, nine latent variables π, η, x, y, z, ψ, ϕ, Ω and Ψ are need to be sampled. However, π, η, ψ, ϕ, Ω and Ψ can be integrated out due to the conjugate priors α, β, γ, δ, τ and μ by using the technique of collapsed Gibbs sampling [40]. Consequently, we only have to sample the community assignment for each user and each document, topic assignment for each document from their conditional distribution given the remaining variables.

First, we need to calculating the joint posterior distribution of DTCD:

$$\begin{aligned}
&p(w, f, t, t', x, y, z|\alpha, \beta, \gamma, \delta, \tau, \mu)\\
&\quad = p(x, y|\alpha)p(w|z, \delta)p(z|y, \gamma)p(f|x, \beta)p(t'|z, \mu)\\
&\quad p(t|x, \tau), \qquad (1)
\end{aligned}$$

where $p(w|z)$ is the probability of the word w generated by the topic z; $p(z|y, \gamma)$ is the probability of the topic z generated by the community y; $p(f|x, \beta)$ is the probability of the user f generated by the community x; $p(t'|z, \mu)$ is the probability of the time stamp t' generated by the topic z;

At each iteration of our Gibbs sampler, DTCD samples both the corresponding community indicator y_{ud} and the topic indicator z_{ud} for each document D_{ud} generated by user u. DTCD samples the corresponding community indicators x_{ui} for each neighbor user i. The sampling formulas are given as follows.

Sampling community indicator x_{ui} for the *i-th* neighbor of user u according to,

$$p(x_{ui} = c* | x_{\neg ui}, f, t, \alpha, \beta, \tau) \propto \frac{n^{(c*)}_{u(f),\neg ui} + n^{(c*)}_{u(d)} + \alpha}{n^{(\cdot)}_{u(f),\neg ui} + n^{(\cdot)}_{u(d)} + |C|\alpha} \cdot \frac{n^{(u)}_{c*(f),\neg ui} + \beta}{n^{(\cdot)}_{c*(f),\neg ui} + |U|\beta} \cdot \frac{n^{(t)}_{c*(f),\neg ui} + \tau}{n^{(\cdot)}_{c*(f),\neg ui} + |T|\tau} \quad (2)$$

where $n^{(c*)}_{u(f),\neg ui}$ and $n^{(\cdot)}_{u(f),\neg ui}$ denote the number of neighbors of user u assigned to community $c*$ and any community with node ui excluded respectively, $n^{(c*)}_{u(d)}$ and $n^{(\cdot)}_{u(d)}$ denote the number of documents of user u assigned to community $c*$ and any community respectively, $n^{(u)}_{c*(f),\neg ui}$ and $n^{(\cdot)}_{c*(f),\neg ui}$ denote the number of user u and any user generated by community $c*$ with node ui excluded respectively, $n^{(t)}_{c*(f),\neg ui}$ and $n^{(\cdot)}_{c*(f),\neg ui}$ denote the number of times that time stamp t and any times tamp of users is generated by community $c*$ with node ui excluded respectively.

Sampling community indicator y_{ud} for the *d-th* document associated with user u according to,

$$p(y_{ud} = c* | y_{\neg ud}, x, z, \alpha, \gamma) \propto \frac{n^{(c*)}_{u(f)} + n^{(c*)}_{u(d),\neg ud} + \alpha}{n^{(\cdot)}_{u(f)} + n^{(\cdot)}_{u(d),\neg ud} + |C|\alpha} \cdot \frac{n^{(z)}_{c*(d),\neg ud} + \gamma}{n^{(\cdot)}_{c*(d),\neg ud} + |Z|\gamma} \quad (3)$$

where $n^{(z)}_{c*(d),\neg ud}, n^{(\cdot)}_{c*(d),\neg ud}$ denotes the number of documents assigned to community $c*$ and generated by topic z and any topic with document ud excluded, respectively.

Sampling topic indicator z_{ud} for the *d-th* document associated with user u according to,

$$p(z_{ud} = z* | y_{ud} = c*, t, z_{\neg ud}, \gamma, \delta, \mu) \propto \frac{n^{(z*)}_{c*,\neg ud} + \gamma}{n^{(\cdot)}_{c*,\neg ud} + |Z|\gamma} \cdot \frac{n^{(t)}_{z*,\neg ud} + \mu}{n^{(\cdot)}_{z*,\neg ud} + |T|\mu} \cdot \frac{\prod_{w=1}^{|W|} \prod_{i=0}^{n^{(w)}_{ud}-1} (n^{(w)}_{z,\neg ud} + \delta + i)}{\prod_{i=0}^{n^{\cdot}_{ud}-1} (n^{(\cdot)}_{z,\neg ud} + |W|\delta + i)} \quad (4)$$

where $n^{(t)}_{z*,\neg ud}, n^{(\cdot)}_{z*,\neg ud}$ denote the number of times that time stamp t and any time stamp of documents is generated by topic $z*$ with document ud excluded respectively, $n^{(w)}_{z,\neg ud}$, $n^{(\cdot)}_{z,\neg ud}$ denote the number of word w and any word is generated by topic $z*$ with document ud excluded respectively. After sampling the community indicator of the neighbor of the user, and the community indicator and topic indicator of the document associated with the user, then we update the counter representing the number of time stamp generated by community which is sampled by formula 2 and the counter representing the number of time stamp generated by topic which is sampled by formula 4. Algorithm 1 describes Gibbs sampling inference for DTCD model.

After a enough number of iterations, the unknown latent variables are calculated follows:

Algorithm 1 Inference on DTCD

Require:
User set U, document set D, edge set E, time stamp set T

Ensure:
Topic assignments Z, community assignments C

```
/* Initialization */
for each u ∈ U do
    for each neighbor node i ∈ N_u of user u do
        c* ~ uniform[1, ..., |C|]
        Assign community c* to edge u → i
    end for
    for each document d associated with user u do
        c* ~ uniform[1, ..., |C|]
        z* ~ uniform[1, ..., |Z|]
        Assign community c* and topic z* to document d
    end for
end for
/* Burn-in*/
I ← number of iterations
i ← 0
while i<I do
    for each user u do
        for each neighbor i of user u do
            Sample a community label c* according to Eq. (2)
            Assign community c* to edge u → i
        end for
        for each document d associated with user u do
            Sample a community label c* according to Eq.(3)
            Assign community c* to document d
            Sample a topic label z* according to Eq. (4)
            Assign topic z* to document d
        end for
    end for
    for each topic z ∈ Z do
        Update Ψ_z
    end for
    for each community c ∈ C do
        Update Ω_c
    end for
end while
```

$$\pi_{u,c} = \frac{n^{(c)}_{u(f)} + n^{(c)}_{u(d)} + \alpha}{n^{(\cdot)}_{u(f)} + n^{(\cdot)}_{u(d)} + |C|\alpha}, \eta_{c,u} = \frac{n^{(u)}_{c(f)} + \beta}{n^{(\cdot)}_{c(f)} + |U|\beta}, \psi_{c,z} = \frac{n^{(z)}_{c(d)} + \gamma}{n^{(\cdot)}_{c(d)} + |Z|\gamma},$$
$$\phi_{z,w} = \frac{n^{(w)}_{z} + \delta}{n^{(\cdot)}_{z} + |W|\delta}, \Omega_{c,t} = \frac{n^{(t)}_{c} + \tau}{n^{(\cdot)}_{c} + |T|\tau}, \Psi_{z,t} = \frac{n^{(t)}_{z} + \mu}{n^{(\cdot)}_{z} + |T|\mu}$$

B. TIME COMPLEXITY

We now analyze the time complexity of our inference algorithm. we compute the community assignments and topic assignments of each document of each user, it takes

$O(|D| \times |C| + |Z| \times |W|)$. We sample the community assignments of each neighbor node of each user, it takes $O(|C| \times |E|)$. Let I denote the number of iterations. The time complexity of the whole inference algorithm is $O((|D| \times |C| + |Z| \times |W| + |C| \times |E|) \times I)$.

V. INFERRING PARAMETERS IN SNAPSHOT NETWORK

A. INFERRING COMMUNITIES IN SNAPSHOT NETWORK

We use the parameters π and η obtained in algorithm (1) to calculate the probability of generating neighbors of users in snapshot network via each community. We can obtain the community membership of the users in the snapshot network. To infer community membership of the user in snapshot network, we assume that the community membership proportions of a user equal to the expectation of the community membership proportions of neighbors linked to the user:

$$P(c|i, t_i = t) = \sum_f (P(c|f)P(f|i, t_i = t)) \tag{5}$$

In Eq.(5), we calculate $P(c|f)$ via Bayes' formula based on the result of parameters estimated in DTCD:

$$P(c|f) = \frac{P(c)P(f|c)}{\sum_c P(c)P(f|c)}, \tag{6}$$

where $P(c) = \pi_{i,c}$, $P(f|c) = \eta_{c,f}$. Then the next issue is to calculate $P(f|i, t_i = t)$. $P(f|i, t_i = t)$ is estimated via the empirical distribution of the neighbor nodes of the users in current snapshot network.

$$P(f|i, t_i = t) = \frac{n_{i,t}(f)}{\sum_f n_{i,t}(f)}, \tag{7}$$

where $n_{i,t}(f)$ denotes the number of user f in the neighbors of node i in t-th snapshot network.

B. INFERRING COMMUNITY MEMBERSHIP OF A TOPIC IN SNAPSHOT NETWORK

We use the parameters π and ψ obtained in algorithm (1) to calculate the probability of generating topics assigned to documents associated with the users in snapshot network via each community. So we can obtain the community membership of topic in the snapshot network. To infer community membership of a topic in a snapshot network, we assume that the community membership proportions of a topic equal to the expectation of the community membership proportions of topic of documents associated with the user:

$$P(c|z, t_z = t) = \sum_d (P(c|d)P(d|z, t_z = t)) \tag{8}$$

In Eq.(8), we calculate $P(c|d)$ by Bayes' formula based on the result of parameters estimated in DTCD:

$$P(c|d) = \frac{P(c)P(d|c)}{\sum_c P(c)P(d|c)}, \tag{9}$$

where $P(c) = \pi_{i_d,c}$, $P(d|c) = \psi_{c,z_d}$, i_d denotes the user i with which the document d is associated, and z_d denotes the topic of document d. The next issue is to determine $P(d|z, t_z = t)$. $P(d|z, t_z = t)$ is estimated via the empirical distribution of the topics of the documents associated with the users in current snapshot network.

$$P(d|z, t_z = t) = \frac{n_{z,t}(d)}{\sum_d n_{z,t}(d)}, \tag{10}$$

where $n_{z,t}(d)$ denotes the number of document d with topic z in the documents associated with the user i in t-th snapshot network.

C. INFERRING TOPICS IN SNAPSHOT NETWORK

We use the parameters ψ and ϕ obtained in algorithm (1) to calculate the probability of generating documents with the topic of users in snapshot network via each word. We can obtain the topic membership of word in the snapshot network. To infer topic membership of a word in the snapshot network, it is assumed that the topic membership proportions of a word equal to the expectation of the topic membership proportions of word in documents associated with the user:

$$P(z|w, t_z = t) = \sum_d (P(z|w_d)P(w_d|z, t_z = t)) \tag{11}$$

In Eq.(11), $P(z|w_d)$ can be calculated via Bayes' formula based on the parameters estimated in DTCD:

$$P(z|w_d) = \frac{P(z)P(w_d|z)}{\sum_z P(z)P(w_d|z)}, \tag{12}$$

where $P(z) = \psi_{c_d,z}$, $P(w_d|z) = \phi_{z,w_d}$, c_d denotes the community of document d, and w_d denotes the word w in document d. The next issue is to determine $P(w_d|z, t_z = t)$. $P(w_d|z, t_z = t)$ is estimated by the empirical distribution of the topic of the documents associated with the users in current snapshot network.

$$P(w_d|z, t_z = t) = \frac{n_{z,t}(w_d)}{\sum_d n_{z,t}(w_d)}, \tag{13}$$

where $n_{z,t}(w_d)$ denotes the number of word w in the document d with topic z in t-th snapshot network.

VI. EXPERIMENTS

Experiments are conducted on two real-world datasets to evaluate the community detection and topic extraction performance of the proposed approach. We quantitatively evaluate the model's performance to extract communities and topics.

All experiments are conducted on a PC with Windows 10, a dual core 3.6GHz CPU and 8G memory. The proposed approach is implemented in Python.

A. SET UP

1) DATASETS

Two real-world dynamic social networks are used in our experimental evaluation, namely online forum Reddit and academic cooperation network DBLP.

REDDIT

The Reddit dataset consists of a part of three discussion blocks in *www.reddit.com* from August 25, 2012 to August 31, 2012. The three discussion blocks are *Science*, *Politics*, *Movies*. Each user publishes a post to launch a discussion, which may be commented by the other users by replying the post. On this dataset, the online forum users are taken as the nodes, the posts published or commented by the users are taken as the corresponding node content. If user u replies to user v, then we consider there is an undirected edge between them. We obtain a word dictionary including 5922 words to describe the node content, after removing common stop words and stemming in the node content.

We partition the network into 7 time stamps and each snapshot network includes the interaction between users and the users' publishing contents in corresponding each day. In all, 3080 users who participate in the discussions generate 5236 edges between them. The ground truth communities of users are extracted from the three discussion blocks in which the users take part. Since we extract the network structure and node content from the three discussion blocks, it is assumed that there are three dynamic communities in the Reddit social network. We set the number of community to be 3 in the experiments to evaluate the effectiveness of the methods if need.

DBLP

The DBLP dataset which is a sub-collection of DBLP data[1] contains academic papers published on 11 international main conferences from 2001 to 2011, it includes the fields of "DM&DB", "AI&ML" and "CV&PR". The 11 conferences include CVPR, ICCV, ECCV, NIPS, AAAI, IJCAI, ICML,KDD, ICDE, ICDM, VLDB and SIGMOD. The DBLP dataset extracts network structure based on the co-author relationship between researchers and extracts node content from the titles of papers published by the researchers. We take the researchers as the nodes in the network. If researcher u co-authors with researcher v, then we make an edge between the corresponding nodes in the network. The title of the paper published by a researcher is considered as the content of the corresponding node in the network. We obtain a word dictionary of size 7317 to describe node content, after removing common stop words and stemming in node content. We only select the researchers with no less than 5 papers published in the conferences from 2001 to 2011 into the network. Finally, we obtain 2554 nodes (researchers) and 9963 edges in the entire network.

We partition the dataset into 11 time stamps, and each snapshot network includes the edges between the researchers and the node content in the corresponding year. We consider the field of the conference in which researchers publish papers as the ground truth communities. Since we extract the network structure and node content from the three fields consists of 11 conferences, it is assumed that there are three dynamic communities in the DBLP social network. We set the number of community to be 3 in the experiments to evaluate the effectiveness of the methods if need.

[1] https://dblp.uni-trier.de/

TABLE 2. Comparison of methods in feature and task.

	features			tasks	
	text	link	time	topic ext	comm det
LDA [18]	√			√	
FacetNet [10]		√	√		√
NEIWalk [12]	√	√	√		√
CPD [33]	√	√		√	√
DTCD(Our method)	√	√	√	√	√

2) BASELINES

We compare the proposed DTCD method with several baseline methods. We select four types of baseline methods. First, the method can find topics from node content. Second, the method can find dynamic community based on network structure. Third, the method can find community and topic in static network with node content. Fourth, the method can find community based on network structure and node content in dynamic network.

Table 2 lists the feature of the baseline methods. Method (1) models text and extracts topics. Method (2) aims at modeling dynamics by considering the relationship between consecutive snapshot network. Method (3) detects dynamic communities by integrating text, network structure and time stamp features. Method (4) extracts communities and their topics by integrating text and network structure. At last, we include dynamic topical community detection method, which extracts communities and their topics by integrating text, network and time stamp features.

Latent Dirichlet Allocation (LDA)

LDA [18] defines a generative process for text. In LDA, text is generated by two latent factors. In LDA, a document is considered as consisting of multiple topics. Like in [40], we set hyperparameters
$\alpha = 50/|Z|$ and $\beta = 0.1$. We adopt LDA for topic modeling comparison.

FacetNet

FacetNet [10] detects dynamic communities with temporal smoothness by considering the community label of nodes in last time stamp only based on network structure. We set parameter α=0.9 which is used to balance the cost of snapshot and the cost of the temporal. We adopt FacetNet for dynamic community detection comparison.

NEIWalk

NEIWalk [12] detects dynamic communities based on heterogeneous random walk by integrating network structures and node content. However, it can not detect topics of communities. We adopt the default value of parameters in [12] i.e., trade-off parameters $\alpha = 1/3$, $\beta = 1/3$, $\gamma = 1/3$,

and random walk parameters $l = 100$, $h = 100$. We adopt NEIWalk for dynamic community detection comparison.

COMMUNITY PROFILING AND DETECTION (CPD)

CPD [33] detects communities and their topics by integrating network structure and node content. However, it does not take time stamps information into consideration. We adopt the default value of parameters in [33] provided by the authors, where $\alpha = 50/|Z|$, $\rho = 50/|C|$ and $\beta = 0.1$. We adopt CPD for both dynamic community detection and topic modeling comparison.

DYNAMIC TOPICAL COMMUNITY DETECTION (DTCD)

It is the method proposed in this paper. DTCD detects communities and their topics by integrating network structure, node content and time information. We set $\alpha = 50/|C|$, $\gamma = 50/|Z|$, $\beta = 0.1$, $\delta = 0.1$, $\tau = 0.1$ and $\mu = 0.1$.

B. COMMUNITY DETECTION

Given the ground truth community in the two real-world datasets, we use normalized mutual information (NMI) [41] to evaluate the performance of the methods.

$$NMI(X|Y) = 1 - \frac{H(X|Y) + H(Y|X)}{2}, \tag{14}$$

where X and Y denote two partitions of the network, and $H(X|Y)$ denotes the normalized conditional entropy of a partition X with respect to Y shown in formula (15).

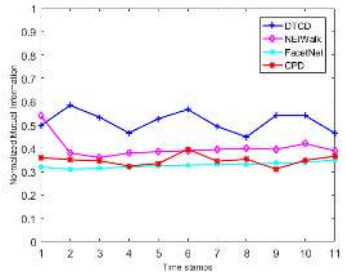
$$H(X|Y) = \frac{1}{|C|} \sum_{k} \frac{H(X_k|Y)}{H(X_k)}, \tag{15}$$

where $|C|$ denotes the number of the community. The larger NMI is, the better the result is. The value of NMI takes from 0 to 1. If it equals to 1 means two partitions match perfectly and equals to 0 on the contrary.

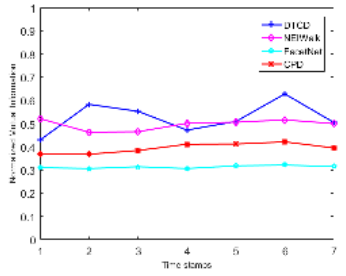
We display the comparison of NMI values in each snapshot network on two datasets in Fig. 3, The NMI values are generated by the four algorithms, namely, DTCD, NEIWalk, Facetnet and CPD. Firstly, it is seen that the methods based on network structure and node content(DTCD, NEIWalk and CPD) achieved better result than the method only base on network structure(Facetnet) does; secondly, our method DTCD achieves the highest NMI on all snapshot network from DBLP and on most of the snapshot network from Reddit. The reason is that topic information contained in the node content helps DTCD to uncover the communities underlying network structures. DTCD does not consider temporal smoothness between consecutive snapshot networks, which leads to that the result achieved by DTCD is more fluctuant.

DTCD finds both communities in each snapshot network and characterize the community temporal variation. After obtaining the community membership of user by inferring communities in snapshot network, we take the index of the maximum value of π_u as the community label of user u.

We can see the different communities with different strength at each time stamp in Fig. 4. It is noticed that each



(a)



(b)

FIGURE 3. The performance of methods on realworld networks. (a) NMI quality comparison on DBLP. (b) NMI quality comparison on Reddit.

TABLE 3. Representative researchers in three communities on the DBLP detected by DTCD.

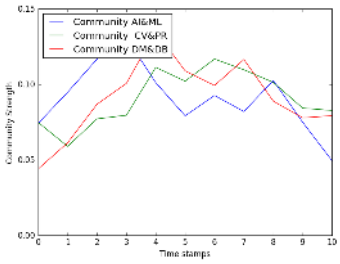
AI&ML	CV&PR	DM&DB
Tuomas Sandolm	Thomas S. Huang	Philip S. Yu
Michael I. Jordan	Larry S. Davis	Jiawei Han
Andrew Y. Ng	Marc Pollefeys	Christos Faloutsos
Bernhar Scholkopf	Luc J. Van Gool	Beng Chin Ooi
Peter Stone	Andrew Zisserman	Jian Pei
Yoshua Bengio	Pascal Fua	Wei Wang
Daphne Koller	Stefano Soatto	Haixun Wang
Vincent Conitzer	Trevor Darrell	Surajit Chaudhuri
Max Welling	Mubarak Shah	Jeffrey Xu Yu
Zoubin Ghahramani	Xiaoou Tang	Qiang Yang

line represents the changing trend of a community in a relative community strength on each time stamp, while there are no relationship between any two communities. In Fig. 4(a), lines represent dome-like shape which describes the real situation. Actually we can obtain that there is more researcher in the middle stage(2004-2009) than preliminary stage (2001-2003) and late stage(2010-2011) from the data. In Fig. 4(b), we can conclude that community 'movies' changes slightly over that 7 days. While community 'politics' and community 'sciences' rise drastically.

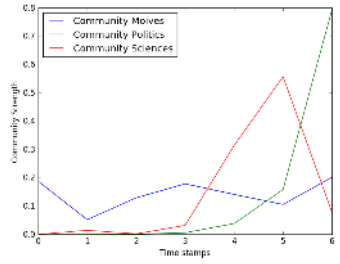
We display the representative researchers of the three communities on the DBLP dataset in Tab. 3. Intuitively, the researchers listed in Tab.3 have some influence in their fields, they also published many papers in the 11 conferences.

C. TOPIC EXTRACTION

Like in [18], we use *perplexity* to evaluate DTCD's topic extraction performance. In topic model, it is a widely used metric. Perplexity measures how well a probability model

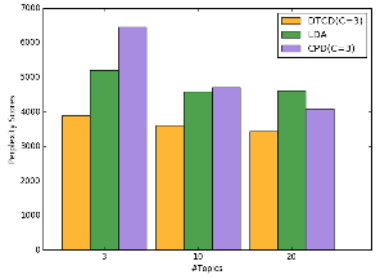


(a)

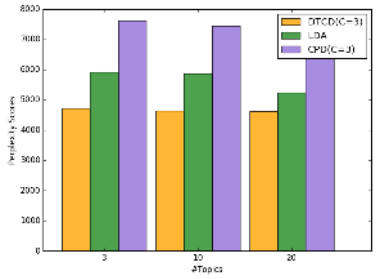


(b)

FIGURE 4. Community temporal variation. (a) Community temporal variation in DBLP. (b) Community temporal variation in Reddit.



(a)



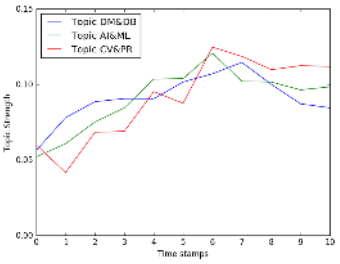
(b)

FIGURE 5. Perplexity scores on realworld networks.(The less, the better). (a) Perplexity scores in DBLP. (b) Perplexity scores in Reddit.

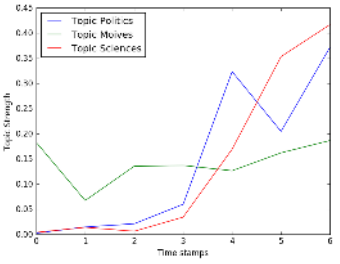
predicts a sample. Perplexity is lower, meaning the model is better. To calculate the perplexity of predicting the test set with M documents, the formula is defined as in formula(16):

$$perplexity(D_{test}) = \exp\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\} \qquad (16)$$

where N_d denotes the number of words in the test document d, and $p(w_d)$ is the probability of the words in the



(a)



(b)

FIGURE 6. Topic temporal variation. (a) Topic temporal variation in DBLP. (b) Topic temporal variation in Reddit.

document; for DTCD, it is calculated as:

$$p(w_d) = \sum_c \pi_{uc} \sum_z \psi_{cz} \prod_l \phi_{zw_{dl}}, \qquad (17)$$

where u is the user with which the document d is associated.

We adopt 5-fold cross validation strategy to calculate the average value of perplexity. In each test, we use 80% of all the documents and all of the edges in network as the train set, and use the remaining 20% of all of documents as the test set.

We conduct some experiments on two datasets as showing results in Fig.5. In Fig. 5(a), it shows the perplexity values with fixed number of community and varying number of topics. We can see that DTCD(|C| = 3,|Z| = 20) achieves the lowest perplexity value which shows that DTCD finds the closest distributions to the real distribution than other methods do. In Fig. 5(b), DTCD(|C| = 3,|Z| = 20) achieves the lowest perplexity value. DTCD assigns a topic to a short document, and gather similar documents via the community, so it can find better topics than LDA. In addition, DTCD takes a series of snapshot network as a whole network with time stamp, while CPD treats the snapshot network as separated one which leads to more perplexity scores than DTCD.

DTCD both find topics and their temporal variations. We can see the different topics with different strength at each time stamp in Fig. 6. It is noticed that each line represents the changing trend of a topic in a relative topic strength on each time stamp, while there are no relationship between any two topics(lines). In Fig. 6(a), lines represent rising trend which describes the real situation. Actually, there are more and more publications on these three topics from 2001 to 2011. In Fig. 6(b), we have conclusions as follows: topic 'movies'

TABLE 4. Representative words of identified topics on DBLP by DTCD.

DM&DB	AI&ML	CV&PR
mining(0.0402)	learning(0.0287)	recognition(0.0405)
patterns(0.0134)	optimal(0.0143)	image(0.0210)
graph(0.0129)	complexity(0.0129)	object(0.0140)
large(0.0127)	convergence(0.0094)	models(0.0131)
frequent(0.0119)	feature(0.0086)	detection(0.0109)
pattern(0.0116)	model(0.0080)	tracking(0.0104)
time(0.0112)	matching(0.0057)	analysis(0.0100)
query(0.0096)	time(0.0055)	motion(0.0088)
indexing(0.0096)	adaptive(0.0055)	shape(0.0082)
tree(0.0080)	online(0.0055)	multiple(0.0072)

TABLE 5. Comparison of the overall calculation time in seconds cost by each method.

Dataset	Reddit	DBLP
FacetNet	834	231
NEIWalk	717	187
CPD	81	118
DTCD	12602	8708

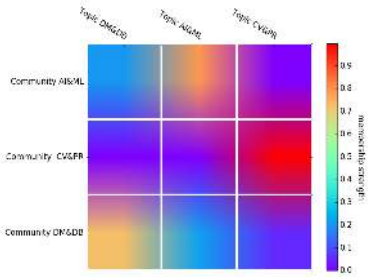
changes slightly over that 7 days; topic 'politics' and topic 'sciences' rise drastically.

In Tab. 4, we illustrate three major topics discovered by DTCD on the DBLP dataset. In each topic, we list examples of the representative words which occur most frequent in the publications. DM&DB topic includes research on data mining and database, such as frequent pattern mining, graph mining, database querying; AI&ML topic includes research on artificial intelligence and machine learning, such as adaptive online learning, bayesian network model learning; and CV &PR topic include research in the fields of pattern recognition and computer vision, such as face recognition, hand motion recognition, contour tracking.

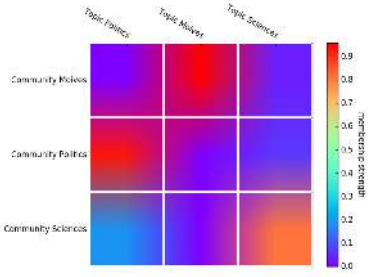
D. COMMUNITY DISTRIBUTION OVER TOPICS

DTCD assumes a community associated with a mixture of topics, instead of a topic. DTCD obtains community distribution over topics via parameter ψ, whose entry $\psi_{c,z}$ denotes the probability of user u belong to community c discussing topic z. In Fig.7, we give the relationship between communities and topics on DBLP and Reddit datasets. We can observe that, the researcher in community 'CV&PR' only focuses on topic 'CV&PR'; the researchers in community 'AI&ML' mainly focus on topic 'AI&ML' and few researchers focus on topic 'DM&DB'; the researchers in community 'DM&DB' mainly focus on topic 'DM&DB' and few researcher focus on topic 'AI&ML'.

As described in section V-C, we can infer topic temporal variation specific on the given community via parameters π and ψ. In Fig.8, we plot the topics temporal variation in community 'AI&ML'. We can observe that the researchers in community 'AI&ML' constantly focuses on topic 'AI&ML' and put little attention on topic 'DM&DB'.



(a)



(b)

FIGURE 7. Community distribution on topic. (a) Community distribution on topic in DBLP. (b) Community distribution on topic in Reddit.

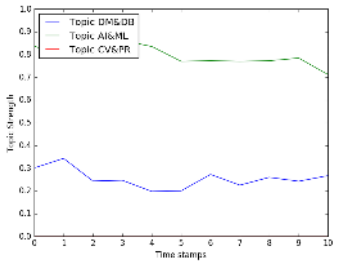


FIGURE 8. Community 'AI&ML' distribution on topic over times.

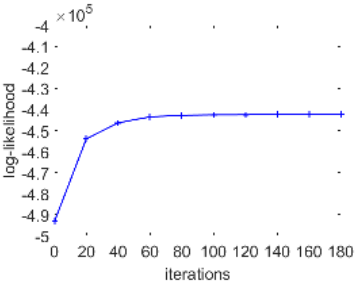


FIGURE 9. The convergence of collapsed Gibbs sampling on DBLP.

It is worthwhile to mention the running time of each method shown in Tab.5. DTCD takes the longest running time. Because it not only detects the communities, but also finds the topics. In additional, DTCD characterizes the community's and the topic's temporal variations over times. In future work, we will study how to improve the efficiency of inference process of DTCD.

Like in [40], we monitor the convergence of the inference algorithm by periodically computing the log-likelihood of observed data, for example, Figure 9 shows the convergence progress of collapsed Gibbs sampling on DBLP.

VII. CONCLUSION

In this paper the problem of dynamic topical community detection is defined. We presented a DTCD model which integrates network structure, text and time into a unified model. DTCD models the community and topic as latent variables, and models the temporal variations of community and topic as multinomial distribution over time stamp. By inferring the latent variables, DTCD can find topic, community and their temporal variations.

We also conducted experiments on two real-world datasets and performed dynamic topical community detection. DTCD outperforms other comparison methods in community detection and topic extraction tasks. DTCD also finds temporal variations of communities and topics. At last we display the relationship between community and topic.

The dynamic topical community detection is a new perspective for dynamic community detection, and there are still several promising future directions. For example, it is interesting to study the mechanism behind the community evolving and the relationship between community evolving and topic changing; From dynamic topical community, it is interesting to find topical influential users in the community which could be applied to viral marketing; another direction is to improve the efficiency of inference process in DTCD; and it is necessary to study the relationship between the different partition of data set and the temporal variations of community and topic. In additional, we can apply DTCD to profile institutions, rank the community by using influence analysis, predict customer loss etc.

REFERENCES

[1] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.

[2] Z. Hu, J. Yao, B. Cui, and E. Xing, "Community level diffusion extraction," in *Proc. SIGMOD*, Melbourne, VIC, Australia, 2015, pp. 1555–1569.

[3] T. Zhu, B. Wang, B. Wu, and C. Zhu, "Maximizing the spread of influence ranking in social networks," *Inf. Sci.*, vol. 278, pp. 535–544, Sep. 2014.

[4] Y. Liu, B. Wang, B. Wu, X. Zeng, J. Shi, and Y. Zhang, "CoGrec: A community-oriented group recommendation framework," in *Proc. Int. Conf. Pioneering Comput. Scientists, Eng. Educators*, Harbin, China, 2016, pp. 258–271.

[5] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks," *Proc. Nat. Aca. Sci. USA*, vol. 101, pp. 5249–5253, Apr. 2004.

[6] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: Membership, growth, and evolution," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, 2006, pp. 44–54.

[7] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "GraphScope: Parameter-free mining of large time-evolving graphs," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, San Jose, CA, USA, 2007, pp. 687–696.

[8] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, Apr. 2007.

[9] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, "Community evolution in dynamic multi-mode networks," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, Las Vegas, NV, USA, 2008, pp. 677–685.

[10] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "FacetNet: A framework for analyzing communities and their evolutions in dynamic networks," in *Proc. Int. World Wide Web Conf.*, Beijing, China, 2008, pp. 685–694

[11] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 4, pp. 16:1–16:36, Nov. 2009.

[12] C.-D. Wang, J.-H. Lai, and P. S. Yu, "NEIWalk: Community discovery in dynamic content-based networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1734–1748, Jul. 2014.

[13] S. Y. Bhat and M. Abulaish, "HOCtracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 1013–1019, Apr. 2015.

[14] Y. Hu, B. Yang, and C. Lv, "A local dynamic method for tracking communities and their evolution in dynamic networks," *Know. Based Syst.*, vol. 110, pp. 176–190, Oct. 2016.

[15] X. Ma and D. Dong, "Evolutionary nonnegative matrix factorization algorithms for community detection in dynamic networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1045–1058, May 2017.

[16] X. Zhou, Y. Liu, B. Li, and H. Li, "A multiobjective discrete cuckoo search algorithm for community detection in dynamic networks," *Soft Comput.*, vol. 21, no. 22, pp. 6641–6652, Nov. 2017.

[17] J. Cheng, X. Wu, M. Zhou, S. Gao, Z. Huang, and C. Liu, "A novel method for detecting new overlapping community in complex evolving networks," *IEEE Trans. Syst., Man, Cyber., Syst.*, to be published.

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

[19] L. Yan, N.-M. Alexandru, and G. Wojciech, "Topic-link LDA: Joint models of topic and author community," in *Proc. Annu. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 665–672.

[20] Y. Zhu, X. Yan, L. Getoor, and C. Moore, "Scalable text and link analysis with mixed-topic link models," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, 2013, pp. 473–481.

[21] H. Alvari, A. Hajibagheri, and G. Sukthankar, "Community detection in dynamic social networks: A game-theoretic approach," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Beijing, China, Aug. 2014, pp. 101–107.

[22] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, Paris, France, 2009, pp. 927–935.

[23] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 718–729, Aug. 2009.

[24] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Community detection with edge content in social media networks," in *Proc. 28th Int. Conf. Data Eng.*, Washington, DC, USA, Apr. 2012, pp. 534–545.

[25] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam, "Using content and interactions for discovering communities in social networks," in *Proc. Int. World Wide Web Conf.*, Lyon, France, 2012, pp. 331–340.

[26] Y. Hu and B. Yang, "Enhanced link clustering with observations on ground truth to discover social circles," *Knowl. Based Syst.*, vol. 73, pp. 227–235, Jan. 2015.

[27] L. Liu, L. Xu, Z. Wangy, and E. Chen, "Community detection based on structure and content: A content propagation perspective," in *Proc. IEEE Int. Conf. Data Mining*, Atlantic, NJ, USA, Nov. 2015, pp. 271–280.

[28] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proc. AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 265–271.

[29] D. He, Z. Feng, D. Jin, X. Wang, and W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 116–124.

[30] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 1151–1156.

[31] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proc. Int. World Wide Web Conf.*, Rio de Janeiro, Brazil, 2013, pp. 1089–1098.

[32] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 1, p. 4, 2014.

[33] H. Cai, V. W. Zheng, Z. Fanwei, K. C.-C. Chang, and Z. Huang, "From community detection to community profiling," *Proc. VLDB Endowment*, vol. 10, no. 7, pp. 817–828, Mar. 2017.

[34] S. Zhao, L. Yu, and B. Cheng, "Probabilistic community using link and content for social networks," *IEEE Access*, vol. 5, pp. 27189–27202, Nov. 2017.

[35] P.-Z. Li, L. Huang, C.-D. Wang, D. Huang, and J.-H. Lai, "Community detection using attribute homogenous Motif," *IEEE Access*, vol. 6, pp. 47707–47716, Aug. 2018.
[36] D.-Y. Nan, W. Yu, X. Liu, Y.-P. Zhang, and W.-D. Dai, "A framework of community detection based on individual labels in attribute networks," *Phys. A, Stat. Mech. Appl.*, vol. 512, pp. 523–536, Dec. 2018.
[37] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, p. 43, 2013.
[38] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, 2006, pp. 424–433.
[39] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Jeju Island, Korea, 2012, pp. 536–544.
[40] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.
[41] S. Alexander and G. Joydeep, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

YUNLEI ZHANG received the B.E. and M.E. degrees in computer science and technology from Hebei University of Science and Technology and Liaoning University of Technology, in 2005 and 2009, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with the Beijing University of Posts and Telecommunications, since 2014. His research interests include data mining, social computing, and social network analysis.

BIN WU received the B.E. degree from the Beijing University of Posts and Telecommunications, in 1991, and the M.E. and Ph.D. degrees from the ICT of Chinese Academic of Sciences, in 1998 and 2002, respectively. He joined the Beijing University of Posts and Telecommunications as a Lecturer, in 2002, where he is currently a Professor. He has published more than 100 papers in refereed journals and conferences. His current research interests include social computing, data mining, and complex network.

NIANWEN NING is currently pursuing the Ph.D. degree in computer science and technology with the Beijing University of Posts and Telecommunications, since 2017. His research interests include data mining, network embedding, and social network analysis.

CHENGUANG SONG received the B.E. and M.E. degrees in electronic and information engineering, and electronic and communication engineering from Xuchang University and Chongqing University of Posts and Telecommunications, in 2014 and 2018, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with the Beijing University of Posts and Telecommunications, since 2018. His research interests include data mining, complex network, and social network analysis.

JINNA LV received the B.E. and M.E. degrees from Zhengzhou University, Zhengzhou, China, in 2006 and 2009, respectively. She is currently pursuing the Ph.D. degree in computer science and technology with the Beijing University of Posts and Telecommunications, since 2015. Her research interests include multimedia content analysis, social relation extraction, and social network analysis.