IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 October 2023, accepted 25 October 2023, date of publication 30 October 2023, date of current version 2 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328616

RESEARCH ARTICLE

Detecting Community Evolution by Utilizing Individual Temporal Semantics in Social Networks

FENG WANG, DINGBO HOU, AND HAO YAN
School of Intelligent Computing Engineering, Tianjin Renai College, Tianjin 301636, China

Corresponding author: Feng Wang (101060@tjrac.edu.cn)

This work was supported by the Tianjin Research Innovation Project for Postgraduate Students under Grant 2021KJ083.

ABSTRACT Social networks are becoming increasingly popular and significant. One of the most distinctive features of these networks is their dynamic nature, which means that they change over time. Consequently, the community structure on these platforms also changes with time, making the detection of community structure a crucial area of research. Specifically, there is still a lack of understanding of how social networks and communities evolve over time. In this paper, we reveal that individual changing topics (i.e., individual temporal semantics) are a vital factor that drives community evolution. A novel dynamic community detection model is proposed, which takes into account natural evolutionary features. The model first partitions social networks into snapshots. It then detects the community structure at each snapshot by utilizing individual changing topics and information from the previous snapshot. Finally, the evolution of users' interested topic distributions and topic distributions of communities are identified. The model is compared with five state-of-the-art baselines on two real datasets, and the experimental results demonstrate that our model outperforms all baselines.

INDEX TERMS Community detection, social network, graphical model, community evolution, topic model.

I. INTRODUCTION

Community detection has garnered considerable attention in the realm of social network analysis [1], [2], [3]. Communities refer to groups of nodes that exhibit strong interconnections within themselves while maintaining sparse connections with nodes from other groups [4]. Understanding the community structure holds crucial importance in comprehending the intricacies of social networks. Nevertheless, social networks evolve over time, leading to continuous changes in their community structure. Conventional static community detection methods fall short in capturing temporal dynamics. Therefore, there is a pressing need to address the challenges of detecting community structures dynamically and modeling the evolving communities in temporal social networks.

Numerous dynamic community detection models have been proposed in the literature [5], [6]. These models approach the analysis of social networks by either partitioning them into successive snapshots or treating them as temporal networks [7], [8]. Three classes of dynamic community detection models have been proposed by researchers. The first class consists of independent community detection methods that partition the network into snapshots and identify communities in each snapshot separately. However, these methods do not consider the connections and relationships between the snapshots. On the other hand, earlier research has mainly focused on the network's topology, neglecting the importance of its content. However, network content plays a crucial role in community detection. The challenge lies in seamlessly integrating both the network's topology and content, a task that has been identified as a significant obstacle [9].

This work is driven by four unsolved challenges in current dynamic community detection models, highlighting the need for further research.

1) The first common drawback of the existing methods is that none of them address the issue of identifying the

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

driving factors behind the evolution of communities. They fail to explain why communities undergo changes over time. This issue holds significant importance for several reasons. First, social networks undergo transformations due to key factors, and understanding these factors is essential for effectively modeling the dynamic nature of social networks and their evolving community structures. Second, the underlying mechanisms responsible for community generation remain unknown, and resolving these issues would enable more accurate community detection and unveil the patterns of community evolution. Moreover, beyond just the community structure, comprehending community semantics is crucial. In dynamic social networks, the meaning and context of communities also change alongside their structures. Therefore, the changing network content plays a vital role in driving the evolution of communities [10].

2) Furthermore, gaining a comprehensive understanding of social networks necessitates research at both the individual and network levels. However, focusing solely on the network scale overlooks the individual-level effects that influence network characteristics. Therefore, it is crucial to consider both aspects appropriately. Communities offer an appropriate research granularity since they are composed of individuals. The shared interests and social behaviors among individuals give rise to community semantics and interactions across different communities. Nevertheless, the process of how communities emerge from an individual perspective remains unknown. The concept of community structure implies a trade-off between network topology and semantics. Even though individuals within a community may communicate more frequently, they might engage in diverse discussions due to different interests [11].

3) Existing studies on dynamic social networks lead us to the conclusion that individual topic shifts play a key role in motivating changes within social networks and communities [12]. As depicted in Fig. 1, when nodes in a network maintain their interests unchanged, the network remains stable. However, if some users alter their topics of interest, they tend to interact with others who share similar interests, leading to the formation of new edges in the network. Therefore, the crucial factor driving community topology and semantics is the individual's changing topics.

4) Moreover, given that community structure involves high-order relationships, certain nodes within the same community may not have direct links [13]. In light of this, a new structure called the ''semantic sub-community'' should be considered. Individuals within a semantic sub-community belong to the same overarching community while also sharing a common topic preference. Therefore, capturing the finer nuances of community dynamics and their underlying semantic associations is significant.

To resolve the above four challenges, we propose a novel community detection model called DCEITS (Detecting Community Evolution by utilizing Individual Temporal Semantics) by seamlessly integrating individual changing topics with network topology and content. DCEITS consists of two key sub-models that address network topology and content, both originating from individual topic changes. During the generation process, individuals form semantic sub-communities based on their discussions and topics. Each edge in the network is evaluated to determine whether it belongs to an intra-community or inter-community relationship, thereby reflecting the connections between different communities. Research has demonstrated the existence of various types of diffusion between communities [14], making it crucial to accurately discriminate these edges to achieve better community detection results.

The contributions and innovations of this research include the following aspects:

1. We unveil the primary drivers behind community evolution, namely, individuals' changing topics, which play a pivotal role in the dynamics of communities over time. This represents the primary innovation in our study.

2. To address the conflict arising from the definition of community, we delve into community semantics, encompassing a resolution between individual-level granularity and community-level granularity, which marks our second innovative contribution. Moreover, we identify node topic distributions and topic distributions of communities.

3. We rigorously evaluate the effectiveness of our proposed model on two real datasets through comprehensive experiments. The results demonstrate that DCEITS outperforms all the baseline models.

The structure of the paper is organized as follows: Section II provides a review of related works in the field. In Section III, we present the intricate details of our proposed model. Section IV outlines the process of model inference. In Section V, we conduct experiments to validate our model's performance. Finally, Section VI concludes the paper and discusses potential future directions for further research.

II. RELATED WORK

Extensive research has been conducted on community detection models [5], [6]. Certain methods focus on static networks and solely aim to identify the static community structure [4]. In such approaches, nodes within networks are grouped, and the communities remain unchanged over time. Recently, there has been significant interest in dynamic community detection techniques due to the ever-changing nature of networks influenced by user activities such as posting and replying. As communities evolve with their members and semantic properties, researchers have focused on addressing this dynamic nature [15], [16].

In dynamic community detection models, networks are divided into snapshots, typically by day, month, or year. By considering the relationships between these snapshots, researchers can detect how communities evolve over time [17]. However, determining the nature of these relationships can be challenging. One approach, presented in [18], involves quantitative analysis of community evolution in dynamic networks. This analysis includes events such as community growth, merging, birth, contraction, splitting, and death.

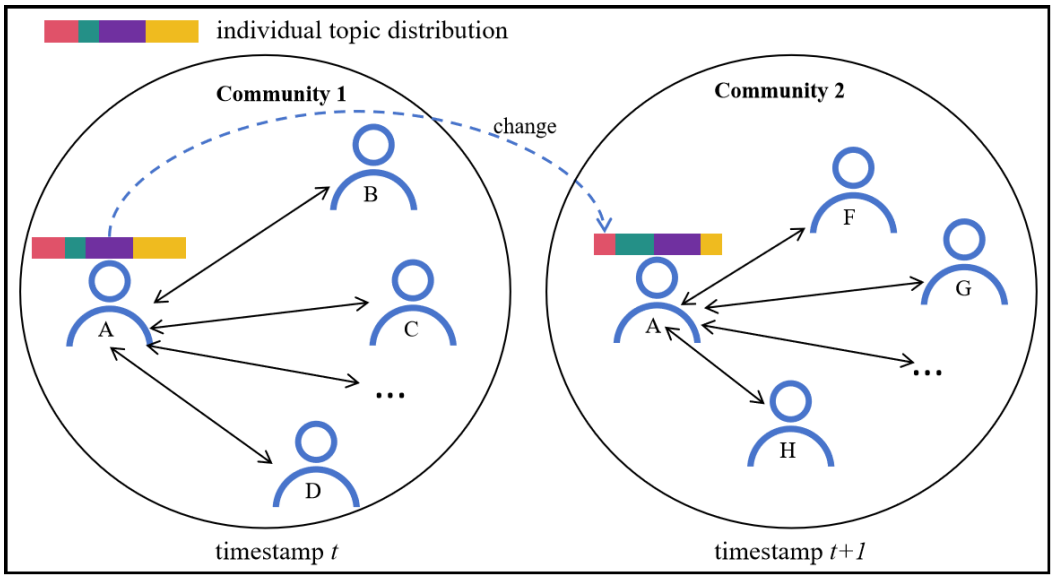


FIGURE 1. A toy social network. If user "A" changes interested topics at some time point, he/she would communicate with users who are also interested in these topics. Then, new edges are generated between these pairs of users and communities also change.

However, the evolution of communities remains a complex procedure, especially the evolution of community semantics. Some methods extend the classical Louvain model [19] and use loss functions to minimize differences between communities in adjacent snapshots. The main objective is to improve efficiency. Another classical approach, introduced in [20], is called the FacetNet dynamic community detection model, which leverages historical community information to analyze community evolution through a unified process.

To intuitively and rigorously integrate community structure and community evolution, a random block model that changes with time was proposed. This model uses a probability transition matrix to store the communities to which all nodes in the snapshot belong [21], [22]. The process involves placing a new node and its newly added or deleted neighbor nodes or edges into a separate community. Another approach in [9] transforms the content network into an edge network addressing the evolving nature of connections and content. The adaptability of the method makes it a valuable contender in understanding complex network structures over time. In [23], community evolution is modeled as a multiple-object optimization problem, resulting in a method called DYNMOGA for dynamic community detection. It is a promising method for understanding how communities evolve over time, providing valuable insights into complex network dynamics. The multiobjective aspect enhances its versatility, making it a valuable tool for researchers in network analysis.

In [24], each community is viewed as a group of follower nodes following potential leader nodes, with a popular node acting as the central node. This innovative method offers a unique perspective on how influential nodes drive community evolution over time. It could be a valuable contribution to understanding dynamic network structures and their key drivers. The concept of closed triplets, proposed in [25], is believed to affect the formation and evolution of networks. This work presents a novel approach to dynamic network embedding by explicitly modeling the triadic closure process. The authors argue that existing network embedding methods often overlook the temporal dependencies and the importance of triadic closure in dynamic networks.

Numerous studies have been proposed to mine community semantics [26], [27], [28], [29]. In the context of community, the discussions among community members are regarded as community semantics. Therefore, topic models are integrated into community detection models. One innovative method, as proposed in [30], takes a holistic approach to dissecting and modeling concise texts within social networks, considering both their spatial and temporal dimensions. This methodology confronts the inherent challenge of comprehending succinct texts in social platforms, such as tweets or status updates, which often lack extensive length and context. Additionally, another study [31] introduces an inventive topic modeling technique that gains insights into user preferences and intentions within social networks. This research addresses the challenge of analyzing large-scale social network data and extracting meaningful information from user-generated content. Moreover, the exploration of dynamic topic modeling finds its niche in the work of [32], which is specifically tailored for short text streams. This approach tackles the challenge of capturing the evolution of topics in short texts, a common phenomenon in social media platforms such as Twitter. Last, a unique approach to uncovering trending topics within vast social network data emerges in [33]. A blend of sparse representation and recurrent neural networks (RNNs) is employed to identify bursty topics. This innovation addresses the distinctive challenge posed by the

TABLE 1. Notations.

Notations	Descriptions
K	set of topics
T	set of snapshots
D	set of posts
E	set of links
W	word set of vocabulary
π	community distribution
θ	topic distribution of communities
ϕ	word distribution of topics k
ψ	network snapshot distribution of community and topic
$\alpha, \beta, \epsilon, \rho$	Dirichlet priors for θ, ϕ, ψ, π

high volume, variety, and speed characterizing social network data.

III. METHODS

A. PROBLEM FORMULATION

All notations are shown in Table 1.

Definition 1: A **social network** with textual content is defined by $G = (U, E, D)$. U, E and D represent sets of users, directed links and documents, respectively.

Definition 2. The **community membership distribution** is defined by π_i, where, i is the user ID and $|C|$ is the number of communities. Element $\pi_{i,c}$ represents the probability of belonging to community c.

Definition 3. A **topic** k is defined by ϕ_k that follows a multinomial distribution over vocabulary. For a word w, the element ϕ_{kw} represents the probability of belonging to topic k. The number of topics is $|K|$.

Definition 4. The **topic distribution of the community at time stamp** t is defined by $\theta_{c,t}$. The element $\theta_{c,t,k}$ represents the probability of belonging to topic k for community c at the time stamp t.

Definition 5. The **time stamp distribution of user and topic** are defined by ψ_{ik}, $i \in U$, $k \in K$, which is a multinomial distribution over network snapshots. $|T|$ is the number of network snapshots.

Definition 6. **Community diffusion** $\eta_{c,c'}$ defines the tendency of forming a link between community c and community c'. When c and c' are closely related, $\eta_{c,c'}$ is larger. Otherwise, the value of $\eta_{c,c'}$ is determined by the correlation of c and c'.

B. MODEL STRUCTURE

In this section, a community detection model that generates network topology and content is designed which is called DCEITS (Detecting Community Evolution by Utilizing Individual Temporal Semantics). Fig. 2 illustrates the probabilistic DCEITS model graphically. It includes two components: a) Network topology component; b) Network content component.

1) NETWORK TOPOLOGY COMPONENT

A link E^t_{ij} from user i to user j is generated as follows. $\eta_{c,c'}$ represents the tendency of forming a link from user i in community c to user j in community c'. Then, a sigmoid function is utilized to generate this link.

$$P(E^t_{ii'} = 1|c_i, c_{i'}, \eta) = \sigma(\eta_{c,c'}) = 1/(1 + e^{-\eta_{c,c'}}) = \frac{1}{2}\int_0^\infty \varphi(\eta_{c,c'}, \xi_{ij})P(\xi_{ij})d\xi_{ij}, \quad (1)$$

where ξ_{ij} is a parameter of the Pólya-Gamma distribution. A joint probability distribution is derived for inference:

$$P(E^t_{ii'} = 1, \xi_{ij}) = \frac{1}{2}\varphi(\eta_{c,c'}, \xi_{ij})P(\xi_{ij}|1, 0) \quad (2)$$

2) NETWORK CONTENT COMPONENT

Network content is generated on the basis of several latent factors, i.e., user's community membership distribution, community-snapshot-topic distribution and topic-word distribution. The content of the current link is generated as follows. User i's community indicator c_{ij} is sampled based on its community distribution π_i, which means that user i belongs to community c_{ij}. Then, the topic z_{iq} is sampled based on community-topic distribution $\theta_{c,t}$ which means that the topic of the current content is z_{iq}. Finally, all words and the snapshot indicator of the content are generated based on topic-word distribution $\phi_{z_{ij}}$ and community-topic distribution over time $\psi_{iz_{ij}}$.

3) GENERATION PROCEDURE

The generation procedure is described as follows.

1) For each snapshot $t = 1, 2, 3 \ldots, |T|$
   a) For community indicator $c = 1, 2, \ldots, C$,
      i) Topic distributions are sampled from $\theta_{ct} \mid \alpha \sim Dir(\alpha)$;
   b) For topic $k = 1, 2, \ldots, K$,
      i) Word distribution is sampled from $\phi_k \mid \beta \sim Dir(\beta)$;
   c) For each user $i = 1, 2, \ldots, U$,
      i) Community distribution is sampled from $\pi_i \mid \rho \sim Dir(\rho)$;
      ii) For community indicator $c = 1, 2, \ldots, C$,
         A) Snapshot distribution is sampled from $\psi_{ic} \mid \epsilon \sim Dir(\epsilon)$
      iii) For each link $q = 1, 2, \ldots,$
         A) Community indicator is sampled from $c_{iq} \mid \pi_i \sim Mul(\pi_i)$;
         B) Sample topic indicator from $z_{iq} \mid \theta_{c_{iq}t} \sim Mul(\theta_{c_{iq}t})$;
         C) Sample the link from i to i': $E^t_{ii'} \mid c_{iq}, c_{i'q}, \eta \sim Ber(\sigma(\eta_{c_{iq},c_{i'q}}))$;
         D) For each word $r = 1, 2, \ldots,$
            - Sample word from $w_{iqr} \mid \phi_{z_{iq}} \sim Mul(\phi_{z_{iq}})$;
         E) Sample snapshot $t_{iq} \mid \psi_{iz_{iq}} \sim Mul(\psi_{iz_{iq}})$;

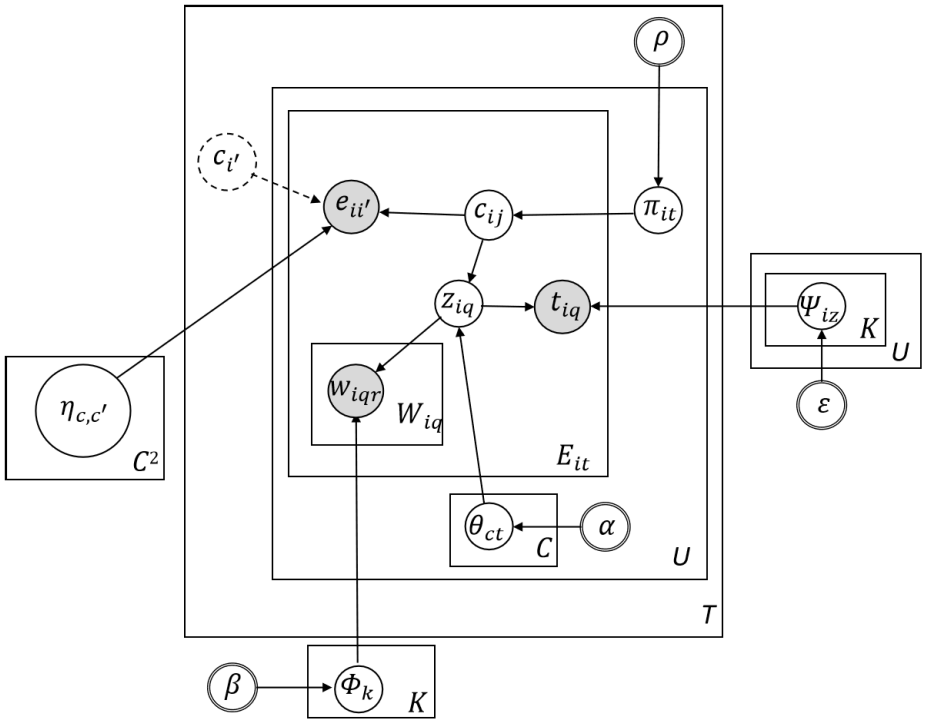


FIGURE 2. The graphical representation of our model.

IV. MODEL INFERENCE

A. APPROXIMATE INFERENCE

In this section, an approximate inference technique is adopted to infer all parameters. The full posterior distribution of the model is:

$$\begin{aligned} &P(\theta,\eta,\psi,\phi,\pi,c,z,\xi|U,E,D,\varepsilon,t,\rho,\alpha,\beta)\\ &\propto P(\theta|\alpha)P(\phi|\beta)P(\pi|\rho)P(\psi|\varepsilon)P(c|\pi)P(z|c,\theta)P(w_d|z,\phi)\\ &\cdot P(t_d|z,\psi)P(e,\xi|I,\eta,c,z). \end{aligned} \tag{3}$$

First, marginalizing $\{\pi,\theta,\phi,\psi\}$ in Eq. (3):

$$\begin{aligned} &P(c,z|.)\\ &\propto \int P(\pi|\rho)P(c|\pi)d\pi \int P(\theta|\alpha)P(z|c,\theta)d\theta\\ &\cdot \int P(\phi|\beta)P(w_d|z,\phi)d\phi \int P(\psi|\varepsilon)P(t_d|c,z,\psi)d\psi\\ &\cdot P(e,\xi)). \end{aligned} \tag{4}$$

The first integral in Eq. (4) is calculated as follows.

$$\begin{aligned} &\int P(\pi|\rho)P(c|\pi)d\pi\\ &= \int (\prod_{i=1}^{|U|}\frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{|C|}}\prod_{c=1}^{|C|}\pi_{ic}^{\rho-1})(\prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{c=1}^{|C|}\pi_{ic}^{n_j^{(c)}})\\ &= \prod_{i=1}^{|U|}\frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{C}}\int \prod_{i=1}^{|U|}\prod_{c=1}^{|C|}\pi_{ic}^{n_i^{(c)}+\rho-1}d\pi\\ &= \prod_{i=1}^{|U|}\frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{|C|}}\frac{\prod_{c=1}^{|C|}\Gamma(n_i^{(c)}+\rho)}{\Gamma(n_i^{(\cdot)}+|C|\rho)}, \end{aligned} \tag{5}$$

where $n_i^{(c)}$ represents the number of times user i is assigned to community c. The dot in $n_i^{(\cdot)}$ denotes the aggregation of all communities. For the second integral in Eq. (4),

$$\begin{aligned} &\int P(\theta|\alpha)P(z|c,\theta)d\theta\\ &= \int (\prod_{c=1}^{|C|}\frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^{K}}\prod_{k=1}^{|K|}\theta_{ctk}^{\alpha-1})\prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{k=1}^{|K|}\theta_{ctk}^{n_{jc}^{(k)}}d\theta\\ &= \prod_{c=1}^{|C|}\frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^{K}}\int \prod_{c=1}^{|C|}\prod_{k=1}^{|K|}\theta_{ctk}^{n_{Dct}^{(k)}+\alpha-1}d\theta\\ &= \prod_{c=1}^{|C|}\frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^{|K|}}\frac{\prod_{k=1}^{|K|}\Gamma(n_{Dct}^{(k)}+\alpha)}{\Gamma(n_{Dct}^{(\cdot)}+|K|\alpha)}, \end{aligned} \tag{6}$$

where $n_{Dct}^{(k)}$ is the number of documents (i.e., contents of links) with topic k in community c at time stamp t. For the third integral in Eq. (4),

$$\begin{aligned} &\int P(\phi|\beta)P(w_d|z,\phi)d\phi\\ &= \int (\prod_{k=1}^{|K|}\frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}}\prod_{w=1}^{|W|}\phi_{kw}^{\beta-1})\prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{w=1}^{|W|}\phi_{zw}^{n_{jz}^{(w)}}d\phi\\ &= \prod_{k=1}^{|K|}\frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}}\int \prod_{k=1}^{|K|}\prod_{w=1}^{|W|}\phi_{kw}^{n_{Dk}^{(w)}+\beta-1}d\phi\\ &= \prod_{k=1}^{|K|}\frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}}\frac{\prod_{w=1}^{|W|}\Gamma(n_{Dk}^{(w)}+\beta)}{\Gamma(n_{Dk}^{(\cdot)}+|W|\beta)}, \end{aligned} \tag{7}$$

where $n_{Dk}^{(w)}$ is the number of times word w in document D is assigned to topic k. The last part in Eq. (4) is calculated by

Eq. (8).

$$\begin{aligned}
&\int P(\psi|\varepsilon)P(t_d|c,z,\psi)d\psi \\
&= \int \prod_{c=1}^{|C|}\prod_{k=1}^{|K|}\frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}}\prod_{t=1}^{|T|}\psi_{ck}^{\varepsilon-1}\prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{t=1}^{|T|}\psi_{cz}^{n_{jcz}^{(t)}}d\psi \\
&= \prod_{c=1}^{|C|}\prod_{k=1}^{|K|}\frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}}\int \prod_{c=1}^{|C|}\prod_{k=1}^{|K|}\prod_{t=1}^{|T|}\psi_{ck}^{n_{Dck}^{(t)}+\varepsilon-1}d\psi \\
&= \prod_{c=1}^{|C|}\prod_{k=1}^{|K|}\frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}}\frac{\prod_{t=1}^{T}\Gamma(n_{Dck}^{(t)}+\varepsilon)}{\Gamma(n_{Dck}^{(\cdot)}+|T|\varepsilon)},
\end{aligned} \tag{8}$$

where $n_{Dck}^{(t)}$ represents the frequency of links associated with topic k in community c being assigned to time stamp t. All latent variables are sampled by Eq. (10) – Eq. (11). Finally, all parameters are calculated by Eq. (12) – Eq. (15).

$$\begin{aligned}
&P(c_{ij}=c|c_{\neg ij},z_{ij}=k,t_{ij}=t) \\
&= \frac{P(c,z)}{P(c_{\neg ij},z)} \\
&= \frac{\int P(\pi|\rho)P(c|\pi)d\pi}{\int P(\pi|\rho)P(c_{\neg ij}|\pi)d\pi}\frac{\int P(\theta|\alpha)P(z|c,\theta)d\theta}{\int P(\theta|\alpha)P(z|c_{\neg ij},\theta)d\theta} \\
&\quad\cdot\frac{\int P(\psi|\varepsilon)P(t_d|c,z,\psi)d\psi}{\int P(\psi|\varepsilon)P(t_d|c_{\neg ij},z,\psi)d\psi} \\
&= \frac{n_{i,\neg ij}^{(c)}+\rho}{n_{i,\neg ij}^{(\cdot)}+|C|\rho}\frac{n_{Dct,\neg ij}^{(k)}+\alpha}{n_{Dct,\neg ij}^{(\cdot)}+|K|\alpha}\frac{n_{Dck,\neg ij}^{(t)}+\varepsilon}{n_{Dck,\neg ij}^{(\cdot)}+|T|\varepsilon}.
\end{aligned} \tag{9}$$

$$\begin{aligned}
&P(z_{ij}=k|z_{\neg ij},c_{ij}=c,t_{ij}=t) \\
&= \frac{P(z,c)}{P(z_{\neg ij},c)} \\
&= \frac{\int P(\theta|\alpha)P(z|c,\theta)d\theta}{\int P(\theta|\alpha)P(z_{\neg ij}|c,\theta)d\theta} \\
&\quad\cdot\frac{\int P(\phi|\beta)P(w_d|z,\phi)d\phi}{\int P(\phi|\beta)P(w_d|z_{\neg ij},\phi)d\phi} \\
&\quad\cdot\frac{\int P(\psi|\varepsilon)P(t_d|c,z,\psi)d\psi}{\int P(\psi|\varepsilon)P(t_d|c,z_{\neg ij},\psi)d\psi} \\
&= \frac{n_{c,\neg ij}^{(k)}+\alpha}{n_{c,\neg ij}^{(\cdot)}+|K|\alpha}\frac{\Pi_{w=1}^{|W|}\Pi_{q=0}^{n_{ij}^{(w)}-1}(n_{Dk,\neg ij}^{(w)}+q+\beta)}{\Pi_{q=0}^{n_{ij}^{(\cdot)}-1}(n_{Dk,\neg ij}^{(\cdot)}+q+\beta)} \\
&\quad\cdot\frac{n_{Dck,\neg ij}^{(t)}+\varepsilon}{n_{Dck,\neg ij}^{(\cdot)}+|T|\varepsilon}.
\end{aligned} \tag{10}$$

$$P(\xi_{ij}|.)\propto e^{-\frac{1}{2}\xi_{ij}\omega_{ij}^2}P(\xi_{ij}|1,0)=PG(1,\omega_{ij}). \tag{11}$$

B. PARAMETER ESTIMATION

$$\hat{\pi}_{ic}=\frac{n_i^{(c)}+\rho}{n_i^{(\cdot)}+|C|\rho}. \tag{12}$$

$$\hat{\theta}_{ctk}=\frac{n_{Dct}^{(k)}+\alpha}{n_{Dct}^{(\cdot)}+|K|\alpha}. \tag{13}$$

$$\hat{\phi}_{kw}=\frac{n_{Dk}^{(w)}+\beta}{n_{Dk}^{(\cdot)}+|W|\beta}. \tag{14}$$

$$\hat{\psi}_{ki,t}=\frac{n_{Dck}^{(t)}+\varepsilon}{n_{Dck}^{(\cdot)}+|T|\varepsilon}. \tag{15}$$

C. ALGORITHM SUMMARIZATION AND TIME COMPLEXITY

Algorithm 1 Inference for DCEITS

Require: U, D, E;
Ensure: $\pi, \theta, \phi, \psi, \eta$;
1: Initialize $\eta, \beta, \rho, \epsilon, \alpha$;
2: **for** $lo = 1 : Ter$ **do**
3: **for** $e \in E$ **do**
4: Sample c_{ij} according to Eq. (10);
5: Sample topic indicator z_{ij} according to Eq. (10);
6: Sample ξ_{ij} according to Eq. (11);
7: **end for**
8: **for** each link $e \in E$ **do**
9: Update η by aggregating community and topic of two endpoint users;
10: **end for**
11: **end for**

Algorithm. 1 outlines the inference procedure. Parameter lo denotes the iterations required for convergence. During steps 4-5, community indicators and topic indicators are sampled. Eq. (10) requires constant time for a specific community. In line 5, the second fraction of Eq. (10) takes $\Theta(|W|)$ for a specific topic. Steps 4-5 have a time complexity of $\Theta(|U|\times|D|\times|C|+|U|\times|D|\times|K|\times|W|)$. In step 6, ξ_{ij} is computed. Eq. (11) takes constant time, therefore, step 6 have a complexity of $\Theta(|E|\times|C|+|E|\times|K|\times|W|)$. Step 9 calculates η, with a time complexity of $\Theta(|E|)$. In summary, the overall complexity is linear with respect to the size of the data.

V. EXPERIMENT

Our model's accuracy for community detection is evaluated on two real datasets compared with five state-of-the-art baselines. All experiments are implemented on a personal computer with Intel 5.3 GHz CPU and 128 GB RAM.

A. DATASETS

To accurately evaluate the community detection results of DCEITS and other baselines, we utilized two authentic datasets with known ground-truth: the Reddit dataset and DBLP dataset, as shown in Table 2.

The Reddit data is sourced from reddit.com and preprocessed according to the methods described in reference [9]. It encompasses four distinct forums: Science, Movie, Olympic Games, and Politics. In this context, users within these forums are treated as nodes, and the connections between posts (replies to other posts) serve as the link contents. We choose a single day as a time snapshot. The dataset is divided into 11 snapshots.

TABLE 2. Summary of datasets with ground-truth.

	#users	#links	#words	#time stamps	#communities	#topics
Reddit	28,932	63,524	13,267	11	4	4
DBLP	62,235	812,422	9,644	13	3	3

TABLE 3. Experimental results comparison on Reddit and DBLP.

	Dataset					
	Reddit			DBLP		
Metric / Method	GNMI	F-score	Jaccard	GNMI	F-score	Jaccard
Louvain	0.36	0.58	0.49	0.22	0.62	0.49
TCCD	0.38	0.61	0.59	0.23	0.69	0.52
DYNMOGA	0.31	0.72	0.55	0.30	0.55	0.59
FacetNet	0.29	0.67	0.51	0.29	0.62	0.48
GHIPT	0.42	0.74	0.61	0.36	0.68	0.59
DCEITS(proposed)	**0.53**	**0.80**	**0.71**	**0.43**	**0.78**	**0.62**

The DBLP dataset represents a network of co-authors in academic publications [36]. For our analysis, we focus on three specific research fields: Image processing, Data mining, and Machine learning, spanning from the year 2011 to 2023, with each year constituting a time snap. In this dataset, individual authors are regarded as nodes, and links are established among authors who collaboratively publish papers.

FIGURE 3. Word clouds for topics: Movie, Politics, Science, and Olympics.

B. BASELINES

Five state-of-the-art baselines are chosen to evaluate our model's accuracy. Some of them model user attributes to detect community structure. They are described as follows:

1) **Louvain** [19]. Louvain is a classical community detection method on static social networks.
2) **Topic Correlations-Based Community Detection (TCCD)** [34]. TCCD is proposed by considering the correlations of different topics in the community detection model.
3) **Dynamic MultiObjective Genetic Algorithms (DYNMOGA)** [23]. Community evolution is modeled by a multiple objects optimization problem.
4) **A Framework for Analyzing Communities and Evolutions in dynamic NETworks (FacetNet)** [20]. This model uses historical community structure information on each network snapshot to analyze community evolution through a unified process.
5) **Community Detection Considering Group Homophily and Individual Personality of Topics (GHIPT)** [35]. It investigates individual personality with regard to topics for community detection.

C. METRICS

We use GNMI (Generalized Normalized Mutual Information), F-score and Jaccard index to evaluate community detection accuracy including overlapping community structure. GNMI is a measure used to quantify the amount of information shared between two sets of data while taking into account the size and distribution of these sets. It is commonly used in clustering and information retrieval tasks to evaluate the quality of clustering algorithms or the performance of feature selection methods. F-score is calculated by: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. The Jaccard index is calculated by: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ (i.e., measuring the similarity of samples A and B).

D. COMPARISON WITH BASELINES

All methods are implemented 10 times. The average value of each metric is recorded. Table 3 shows the result comparisons between the baselines and DCEITS on the two datasets with regard to an average score of all snapshots. The best scores are in bold font.

According to Table 3, we obtain the following observations: (1) Our model achieves 26.19% GNMI improvement, 8.1% F-score improvement and 16.39% Jaccard improvement over the second-best baseline on Reddit. For the DBLP dataset, Table 3 shows that our model achieves 19.44% GNMI improvement, 13.04% F-score improvement and 5.08% Jaccard improvement over the second-best baseline on DBLP. DCEITS outperforms all baselines for all metrics. The main reason is that DCEITS processes topics at the individual level that are integrated together to form the topics at the community level. Moreover, individual topics might change over time, which leads to changes of community structure and community semantics. DCEITS is more capable of processing the above situations. (2) GHIPT is better than other baselines. Because it considers users' characteristics (i.e., whether sharing similarities with others within the same community or not) that leads to diverse homophily rates in social networks. (3) The Louvain method considering only network topology obtains the worst score, which proves

FIGURE 4. Word clouds for topics: Image processing, Data mining, and Machine learning.

FIGURE 5. Word clouds for topics: Covid and SARS.

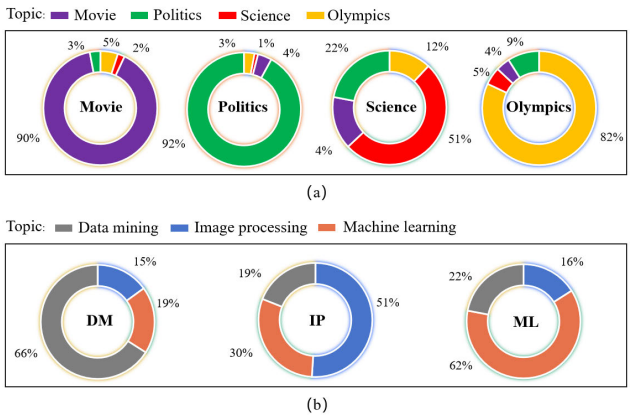


FIGURE 6. Topic distributions of communities.

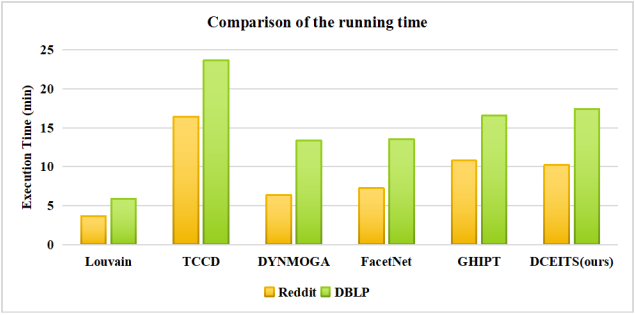


FIGURE 7. Comparison of the running time.

that integrating network topology and network content is significant for community detection.

E. CASE STUDY

In this section, we analyze the word distributions of topics, and the topic distributions of communities. (i.e., $\{\phi, \theta\}$).

1) WORD DISTRIBUTIONS OF TOPICS

We choose the top 20 words in each topic, excluding words with a probability less than 0.001, to generate word cloud illustrating the quality of the topics detected by DCEITS. In these visualizations, word size corresponds to probability, with larger fonts indicating higher probabilities.

In Fig. 3, the word distributions for Movie, Politics, Science, and Olympics topics in the Reddit dataset are displayed. The Movie topic prominently features terms such as "movie," "film," "watch," and "release." The Politics topic includes words such as "politic," "election," "republican," "Obama," and "president." The Science topic focuses on words such as "science," "time," "universe," and "study," while the Olympics topic highlights terms such as "Olympics," "athlete," "London," and "win", as the Olympic Games were held in London in 2012.

Fig. 4 shows the word distributions for Image Processing, Data Mining, and Machine Learning topics in the DBLP dataset. In the Image Processing topic, key research terms include "image," "recognition," "model," and "learn." Data Mining topic is characterized by words like "database," "query," "data," "stream," and "mine." In the Machine Learning topic, important research keywords are "clustering," "classify," "learn," and "model."

The CORD-19 dataset (COVID-19 Open Research Dataset) consists of research papers about COVID-19. A citation network is constructed based on this dataset. The topics detected by DCEITS are demonstrated by word cloud, as shown in Fig. 5. It shows that keywords "COVID," "transmissible," "diseases," and "infectious" are the most important keywords relating to the Covid topic. For the SARS topic, the keywords "coronaviruses," "viruses," "immunized," and "transmissible" receive the greatest attention.

Through the above analysis, we conclude that each topic detected by our model is meaningful.

2) TOPIC DISTRIBUTIONS OF EACH COMMUNITY

In Fig. 6(a), four doughnut charts represent four communities (i.e., Movie, Politics, Science, and Olympics). Each color on the doughnuts denotes one topic. As it shows, the topics of Movie, Politics, and Olympics are dominant in the Movie, Politics, and Olympics communities, respectively. However, for the Science community, although the Science

topic is dominant, 22 percent of posts talk about Politics and 12 percent of posts talk about Olympics.

In Fig. 6(b), three doughnut charts represent three communities (i.e., Data mining, Image processing, and Machine learning). As shown, the topics for Data mining and Machine learning are dominant in the Data mining and Machine learning communities, respectively. However, for the Image processing community, although the Image processing topic is dominant, 30 percent of papers concern Machine learning topic and 19 percent of papers concern Data mining topic due to their interdisciplinary connections.

F. COMPARISON OF THE RUNNING TIME

To compare the execution time between our model and baselines, all methods are implemented 10 times. The average running time is recorded, as shown in Fig. 7. Because the Louvain, FacetNet, and DYNMOGA methods only use network topology, they obtain a low run time. TCCD, GHIPT, and DCEITS all use network topology and network content. TCCD considers topic correlations and isolated posts that are not replied by others. Therefore, TCCD obtains a high running time. Since GHIPT considers invidual characteristics and DCEITS considers individual level topics and both use Gibbs sampling, their running times are comparable.

G. PARAMETER SETTINGS

The sets C and K are of real values. η is randomly initiated. For the Dirichlet hyper-parameters, we run DCEITS under different values. The results show that DCEITS is not sensitive to Dirichlet hyper-parameters, therefore, we set them to predefined values. For the threshold for determining overlapping communities, we test its values from $1/|C|$ to 0.5 with a step of 0.1. The experiments show that $1/|C|$ is the best value. For each user, we choose those communities with probabilities higher than the threshold as its real communities.

VI. CONCLUSION

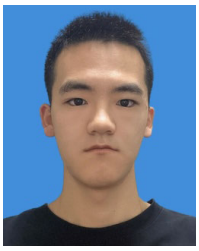
First, we investigated and assessed the influence and importance of considering individual changing topics for dynamic community detection. Individual changing topics reflect users' habits and make significant contributions to the interactions among the users (i.e., the topology structure and contents of a network). Second, we propose a novel method (DCEITS) by combining user changing topics, network topology and network contents uniformly in a generative model. It investigates the formation of a network with complex contents to infer community structure and community topics. Third, we evaluate DCEITS on two real datasets and compare them with five state-of-the-art methods. Experimental results indicate that DCEITS improves the accuracy of community detection. In addition to the accurate identification of community structures, DCEITS can also identify major topics in each community. In the future, we intend to investigate how community topics evolve as a result of the changing of users' interests.

REFERENCES

[1] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.
[2] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surveys*, vol. 45, no. 4, pp. 1–35, Aug. 2013.
[3] V. Satuluri, Y. Wu, X. Zheng, Y. Qian, B. Wichers, Q. Dai, G. M. Tang, J. Jiang, and J. Lin, "SimClusters: Community-based representations for heterogeneous recommendations at Twitter," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3183–3193.
[4] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
[5] P. Wadhwa and M. P. S. Bhatia, "Community detection approaches in real world networks: A survey and classification," *Int. J. Virtual Communities Social Netw.*, vol. 6, no. 1, pp. 35–51, Jan. 2014.
[6] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, Feb. 2010.
[7] X. Su, S. Xue, F. Liu, J. Wu, J. Yang, C. Zhou, W. Hu, C. Paris, S. Nepal, D. Jin, Q. Z. Sheng, and P. S. Yu, "A comprehensive survey on community detection with deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 9, 2022, doi: 10.1109/TNNLS.2021.3137396.
[8] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, P. S. Yu, and W. Zhang, "A survey of community detection approaches: From statistical modeling to deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1149–1170, Feb. 2023.
[9] 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.
[10] Y. Pei, N. Chakraborty, and K. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 2083–2089.
[11] D. Jin, Z. Liu, D. He, B. Gabrys, and K. Musial, "Robust detection of communities with multi-semantics in large attributed networks," in *Proc. Int. Conf. Knowl. Sci., Eng. Manag.* Cham, Switzerland: Springer, 2018, pp. 362–376.
[12] L. Shi, J. Luo, P. Zhang, H. Han, D. El Baz, G. Cheng, and Z. Liang, "Understanding user preferences in location-based social networks via a novel self-attention mechanism," *Sustainability*, vol. 14, no. 24, Dec. 2022, Art. no. 16414.
[13] Y. Zhang, Y. Xiong, Y. Ye, T. Liu, W. Wang, Y. Zhu, and P. S. Yu, "SEAL: Learning heuristics for community detection with generative adversarial networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1103–1113.
[14] J. He, Z. Hu, T. Berg-Kirkpatrick, Y. Huang, and E. P. Xing, "Efficient correlated topic modeling with topic embedding," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 225–233.
[15] R. Márquez, "Overlapping community detection in static and dynamic networks," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 925–926.
[16] T. Li, W. Wang, X. Wu, H. Wu, P. Jiao, and Y. Yu, "Exploring the transition behavior of nodes in temporal networks based on dynamic community detection," *Future Gener. Comput. Syst.*, vol. 107, pp. 458–468, Jun. 2020.
[17] M. G. Bhattacharjee and M. Banerjee, "Change point estimation in a dynamic stochastic block model," *J. Mach. Learn. Res.*, vol. 21, no. 107, pp. 1–59, 2020.
[18] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, Apr. 2007.
[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.
[20] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "Analyzing communities and their evolutions in dynamic social networks," *ACM Trans. Knowl. Discovery From Data*, vol. 3, no. 2, pp. 1–31, Apr. 2009.
[21] N. Mehta, L. Carin, and P. Rai, "Stochastic blockmodels meet graph neural networks," in *Proc. ICML*, vol. 97, 2019, pp. 4466–4474.
[22] X. Wu, P. Jiao, Y. Wang, T. Li, W. Wang, and B. Wang, "Dynamic stochastic block model with scale-free characteristic for temporal complex networks," in *Proc. DASFAA*, vol. 11447, 2019, pp. 502–518.
[23] F. Folino and C. Pizzuti, "An evolutionary multiobjective approach for community discovery in dynamic networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1838–1852, Aug. 2014.

[24] W. Gao, W. Luo, and C. Bu, "Evolutionary community discovery in dynamic networks based on leader nodes," in *Proc. Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2016, pp. 53–60.
[25] L.-K. Zhou, Y. Yang, X. Ren, F. Wu, and Y. Zhuang, "Dynamic network embedding by modeling triadic closure process," in *Proc. AAAI*, 2018, pp. 571–578.
[26] J. Cheng, W. Li, K. Han, Y. Tang, C. He, and N. Zhang, "SARNMF: A community detection method for attributed networks," in *Proc. IEEE 25th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2022, pp. 879–884.
[27] W. Liu, Z. Chang, C. Jia, and Y. Zheng, "A generative node-attribute network model for detecting generalized structure and semantics," *Phys. A, Stat. Mech. Appl.*, vol. 588, Feb. 2022, Art. no. 126557.
[28] V. Moscato and G. Sperlì, "Community detection over feature-rich information networks: An eHealth case study," *Inf. Syst.*, vol. 109, Nov. 2022, Art. no. 102092.
[29] C. He, Y. Zheng, J. Cheng, Y. Tang, G. Chen, and H. Liu, "Semi-supervised overlapping community detection in attributed graph with graph convolutional autoencoder," *Inf. Sci.*, vol. 608, pp. 1464–1479, Aug. 2022.
[30] F. Kou, J. Du, Z. Lin, M. Liang, H. Li, L. Shi, and C. Yang, "A semantic modeling method for social network short text based on spatial and temporal characteristics," *J. Comput. Sci.*, vol. 28, pp. 281–293, Sep. 2018.
[31] L. Shi, G. Song, G. Cheng, and X. Liu, "A user-based aggregation topic model for understanding user's preference and intention in social network," *Neurocomputing*, vol. 413, pp. 1–13, Nov. 2020.
[32] L. Shi, J. Du, M. Liang, and F. Kou, "Dynamic topic modeling via self-aggregation for short text streams," *Peer-Peer Netw. Appl.*, vol. 12, no. 5, pp. 1403–1417, Sep. 2019.
[33] L. Shi, J.-P. Du, M.-Y. Liang, and F.-F. Kou, "SRTM: A sparse RNN-topic model for discovering Bursty topics in big data of social networks," *J. Inf. Sci. Eng.*, vol. 35, no. 4, pp. 749–767, 2019.
[34] Y. Wang, D. Jin, K. Musial, and J. Dang, "Community detection in social networks considering topic correlations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 321–328.
[35] Y. Wang, D. Jin, C. Yang, and J. Dang, "Integrating group homophily and individual personality of topics can better model network communities," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 611–620.
[36] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. 40th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (SIGKDD)*, 2008, pp. 990–998.

FENG WANG received the B.E. degree in precision instrument and the M.E. degree in software engineering from Tianjin University, Tianjin, China, in 1999 and 2014, respectively. She is a Senior Engineer with the School of Computer Science and Technology, Tianjin Renai College. Her current research interests include public safety and the IoT.



DINGBO HOU is currently pursuing the bachelor's degree with the School of Computer Science and Technology, Tianjin Renai College. His current research interests include community testing and deep learning.

HAO YAN is currently pursuing the bachelor's degree with the School of Computer Science and Technology, Tianjin Renai College. His current research interests include community testing and deep learning.