IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 September 2022, accepted 21 September 2022, date of publication 26 September 2022,
date of current version 20 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3209704

RESEARCH ARTICLE

Community Detection in Social Networks Considering Social Behaviors

YINGKUI WANG[1], DI JIN[2], DONGXIAO HE[2], KATARZYNA MUSIAL[3], AND JIANWU DANG[4], (Member, IEEE)

[1]Department of Computer Science and Technology, Tianjin Renai College, Tianjin 300350, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin 300354, China
[3]Faculty of Engineering and IT, School of Software, Advanced Analytics Institute, University of Technology Sydney (UTS), Sydney, NSW 2007, Australia
[4]School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa 923-1292, Japan

Corresponding author: Di Jin (jindi@tju.edu.cn)

This work was supported in part by the Tianjin Research Innovation Project for Postgraduate Students under Grant 2021KJ083; in part by the Natural Science Foundation of China under Grant 61876128, Grant 61772361, and Grant U2031142; and in part by the National Basic Research Program of China under Grant 2013CB329301.

ABSTRACT The study of community detection in networks has drawn great attention in recent years. To find communities and to understand community semantics, both network topology and network content are utilized. Unfortunately, none of them can explain the driving factors of generating community structure with semantics, which is significant for understanding the mechanisms of community generation. Our observations on a large number of networks show that specific user social behaviors are underlying factors for the generation of community structure. We exploit four types of social behaviors that widely exist in networks, i.e., reciprocity of interactions, posting preference, multitopic preference, and temporal variation of topics. We investigate their impacts on the formation process of links and content in networks, during which communities with topics form. Our analysis shows that they are highly related to community structure. Consequently, a generative community detection model SBCD (social behavior-based community detection) is proposed by combining network topology and content, in which the above social behaviors play a core role. The model is evaluated on two real datasets. The experimental results show that SBCD outperforms state-of-the-art baselines. Finally, a case study illustrates several significant observations with respect to the proposed social behaviors.

INDEX TERMS Community detection, social network, graphical model, social behavior.

I. INTRODUCTION

Community detection is one of the hot research topics in the network science field [1], [2], [3], [4]. We can better understand networks and their functions by identifying community structure, which is an inner characteristic of them. A community is defined as a group of nodes that are densely connected but have sparser connections with the nodes from other groups [5]. A large number of community detection methods have been proposed [6], [7], [8], [9], [10], [11], [12], [13], [14]. Some of them utilize only network topology, while others integrate network content into their models. The adoption of network content (e.g., posts, tweets) makes the understanding of community semantics possible [15], [16], [17]. Topics that are discussed in a community are considered as community semantics. Due to the homophily [18], people with similar attributes (interested topics) communicate with each other more frequently and produce denser links, which generates community structure. However, attributes might be inconsistent with topology in the perspective of community structure. Therefore, the attributes of nodes should be processed carefully to improve community detection accuracy, e.g., setting a balance value between topology and attributes to fit the different structure-attributes correlation [19], [20], [21].

Network topology and network content are the outcome of users' social behaviors. One of the elements of users' social behaviors is users' activities [22] (e.g., interactions between

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

users and publishing posts). In social networks, e.g., Twitter, Facebook, and Reddit, users communicate with each other to express ideas by retweeting or commenting on tweets/posts. Relationships between users are generated as links and result in complex network topology. All texts posted by users make up network content that includes semantic information and reflects topics of interest for users. In the generation process of networks, community structure is generated as well.

Considering the relations between network generation and social behaviors, an issue is raised regarding how users' social behaviors affect community structure and community semantics. Investigating the issue is critical to reveal the fundamental generation mechanism of community structure with semantics. To resolve this challenge, there are two key questions that need to be answered.

1) For each link between two users, what is the process of its formation with respect to community structure? Since network topology is the most important information based on which the communities are detected, we consider the formation process of all links. Suppose that user i publishes a tweet about some topics. Later, user j retweets it. Then, a directed link from user j to user i forms. The reasons for the generation of this link are related to two users' intents (i.e., user j is interested in the topics proposed by user i) and latent relationships of two users' attributes (i.e., they are in the same community or they have similar background attributes). In other words, the reasons not only relate to the community distributions of users but also to their topics of interest. Unfortunately, how to model link generation by considering community structure, community semantics and users' social behaviors has not yet been well addressed.

2) How does each user make contributions to the formation of network semantics? This is a critical element regarding the identification of community semantics [23], [24], [25], [26]. Network content is processed in association with network topology to detect community semantics. To answer this question, we need to consider the following two aspects. First, there are two types of content in networks, i.e., link content and node content. For example, a tweet that is not used to respond to others is node content, which means that the tweet is not on a link. A tweet with which a user replies to others is considered as link content. They play different roles in community detection because node content is the first step in creating a link. Therefore, separating link content and node content can accurately model network topology with community structure and network content, which further improves the accuracy of community detection. Moreover, considering the integrity of the dataset, even if a user has never interacted with others but only posted some posts, he should not be deleted from the dataset to better identify community semantics. Second, users in a network might publish posts or have discussions on multiple topics, i.e., users are interested in multiple topics [27]. Moreover, their topics of interest might change with time. How to accurately identify users' changing topics in community detection models is not well resolved.

Although many models have been proposed for community detection and have achieved great improvements, above-stated questions have not been resolved completely when considering social behaviors. Users' social behaviors not only include user activities but also social context (e.g., users' topic interests and posting habits). In this paper, we investigate four types of user social behaviors, which can help answer the above questions. They are stated as follows.

1) **Reciprocity of interactions** represents the tendency of generating a link between two users, e.g., from user i to user j. The probability of this tendency is highly connected with three factors: a) whether those two users belong to the same community, b) whether those two users are interested in the same topics [28], and c) whether target user j is popular or authoritative in a community. This social behavior indicates that users are more likely to communicate with each other when they are both in the same community and have the same topic preferences. They are less likely to have mutual relationships when they are in the same community but with different topic preferences. They are most unlikely to form links when they are neither in the same community nor have the same topic preferences. On the other hand, if user i is a popular user or a person of authority, the probability that others will reply to his or her activity/message is higher. Moreover, as people might have similar interests with people from outside the community, links exist not only inside communities but also between communities. Considering the reciprocity of interactions can explain the fundamental generation process of community structure. This social behavior can resolve the problems caused by the first question, i.e., how links are generated with respect to community structure and community semantics.

2) We take users' **posting preference** into consideration. Take Twitter as an example. When a user sends a tweet, he intends to express his idea on a certain topic. At this time, he cannot predict who will respond. On the other hand, when another user replies to or retweets the tweet with his own words, a link is generated. After analyzing a large number of real networks, we find that each user in a network has his or her own habits. As illustrated in Fig. 1, users can be classified into three categories. a) Users in the first category prefer publishing posts, and they seldom interact (reply or retweet) with others. Some of these posts are never commented on/replied to by others, which is a common case in social networks. Therefore, they do not make any contributions to the forming of links but make contributions only to the formation of network semantics. Some other posts might be commented on very actively by other users. These posts trigger link creation between users. b) Users in the second category like both publishing posts and interacting with others. These users contribute to network semantics with link content and post content. c) In the third category, users always interact with others, but they do not post documents. In this situation, all the contributions they make to the network semantics are link content. In conclusion, users' posting behaviors affect the formation of network semantics in different ways.

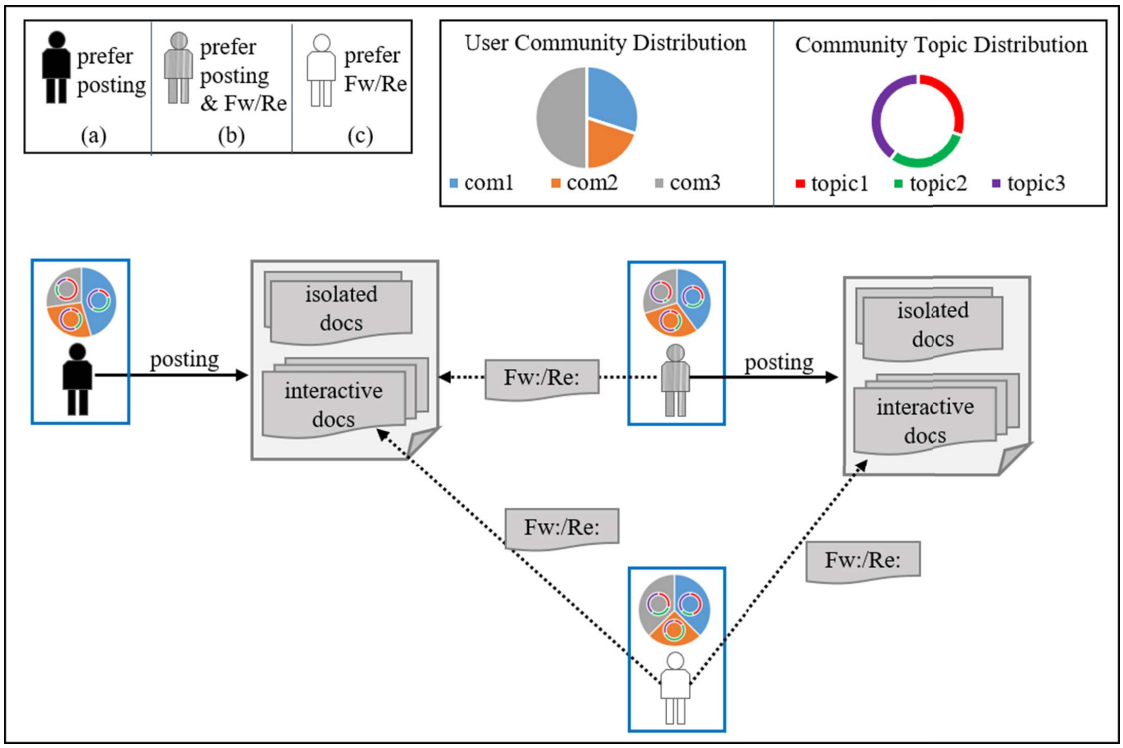

FIGURE 1. **The generation process of a social network with community structure. Users are divided into three categories. a) Users in the first category prefer publishing posts and they seldom interact with others; b). Users in the second category both like publishing posts and interacting with others; c). Users in the third category always interact with others while seldom publish posts. Suppose that there are three communities (i.e., *com1*, *com2*, and *com3*) and three topics (i.e., *topic1*, *topic2*, and *topic3*). Pie charts denote user-community distribution. Doughnut charts denote community-topic distribution. Each user in rectangles has his own community distribution (shown on top of a person). Users post two types of documents, i.e., *isolated docs* that are not replied/forwarded by others and *interactive docs* that are replied/forwarded by others. *Fw:/Re:* denotes replying or forwarding posts of others with text content.**

This behavior can resolve the first aspect of the second question, i.e., separating link content and node content.

3) Users always focus on multiple topics. **Multitopic preference** indicates that users in a network publish posts/tweets focusing on multiple topics. This social behavior has two effects on community structure. First, a user might interact with others who are not in his community, which creates intercommunity links. This increases the complexity of the community structure. Second, users' posts are fundamental factors in generating community semantics. Since a community is a set of users who share similar topics [29], users' interests in multiple topics result in multiple topic distributions in communities. Considering that individual topic distribution results in too many parameters in our model because of the large number of users in networks, in this paper, we consider community-level topic distribution. This behavior can resolve the second aspect of the second question, i.e., users who are interested in multiple topics.

4) Users' topics of interest might change over time. We call this social behavior the **temporal variation of topics**. Individual topic changes lead to the creation of new topical content from a user and new links between users who were not connected previously, which results in changes in network content and network topology. Moreover, observing users' topic-changing patterns is significant when we focus on a specific user, e.g., a leader in a community. This behavior also addresses the second aspect of the second question, i.e., identifying users' changing topics in the community detection model.

In summary, user social behaviors are closely related to the generation of community structure and community semantics. The above four social behaviors give us clues to reveal the mechanism of the generation of community structure. Based on them, we propose a novel generative community detection model to generate network topology, node content and link content. Our contributions are summarized as follows:

- We validate that specific user social behaviors are highly related to the generation of communities. We propose four social behaviors that can accurately describe the inner relations of community structure, community semantics, network topology and network content.
- We propose a novel unified community detection model by integrating network topology, node content, link content and four types of user social behaviors.
- We conduct sufficient experiments to verify our observations of the proposed social behaviors. The results show that our model achieves a more precise community structure than baselines. Moreover, our case studies illustrate the existence of the proposed social behaviors.

The rest of the paper is organized as follows: Section 2 reviews related works; in Section 3, we describe the details of our model; Section 4 describes the model inference; Section 5 shows the experiments and results; finally, in Section 6, the paper is concluded, and future work is presented.

II. RELATED WORK

In this section, related work for community detection is reviewed. We discuss the main contributions of these studies and represent the differences between them and our model.

Many community detection methods have been proposed in recent years, and they have been reviewed in survey articles: [9], [11], [12], [13], which show great interest and a large body of knowledge in this field. Some of them utilize topological information of networks [30], [31]. The goal of these studies is to divide nodes into groups, i.e., communities, without considering any semantic information [32], [33].

Today's social networks exhibit rich information, e.g., user profiles and post content. Utilizing such content to identify semantic information has become a hot topic [34], [35], [36], [37], [38], [39]. A survey [20] is conducted to study attributed community detection task. The work of [40] proposes a comparative study of some existing attributed network community detection algorithms on both synthetic dataset and on real world dataset. A model that generates synthetic node attributed graphs with planted communities is proposed by [41]. The work of [21] proposes a unified weight-based attributed community detection model. Network content reflects individual interested topics. According to homophily principle [18], nodes with similar interested topics are more likely to communicate with each other, which can help improve community detection accuracy [42], [43]. Therefore, an increasing number of studies integrate network topology and network content in community detection models.

To protect the privacy of users, the storage of social data is decentralized, which leads to the definition of Decentralized Online Social Networks (DOSNs) [44]. Several dynamic community detection methods have been proposed in DOSNs [45], [46], [47].

Beyond community structure and community semantics, recent studies have begun to investigate community-level diffusion, i.e., modeling the diffusion patterns of topics across different communities [48] and community profiling [49]. The work in [49] characterizes the intrinsic nature and extrinsic behavior of a community, i.e., community profiling. It considers the heterogeneity among user links (i.e., friendship links and diffusion links). If two users interact with each other to diffuse information, then diffusion links are formed (e.g., a retweet is considered as a diffusion link in Twitter). Both [48] and [49] propose a unified latent framework in which node content and links are both generated by the same latent variables (e.g., community distribution variable, topic distribution variable and word distribution variable). In addition to community detection, they efficiently identify temporal topics within communities, diffusion of communities [48] or community profiling [49].

Investigating social behaviors has been a hot topic in social network research. Social behaviors include the actions of users, e.g., reposting actions [50], commenting on other users' posts, or following others. In [51], the authors investigate the emotional contagion between nodes by considering community structure. The work of [52] classifies users into three roles, i.e., opinion leader, structural hole spanner, and ordinary user. In [53], the authors propose a community detection model based on NMF, in which nodes are considered in the roles of hubs (also known as leaders) or outliers. The work of [54] models user behavior knowledge from different networks to alleviate the data sparsity problem. The work of [55] investigates retweeting behavior on Twitter. It proposes a factor graph model to predict retweeting behavior. The impact of user's availability on on-line ego networks are analyzed by [56]. Communities are detected based on user activity and social ties in [57].

However, existing studies do not consider the underlying factors for community generation. Our model is different from above models. We use four types of social behaviors to generate all posts and links. We show that considering them indeed improves the outcomes of community detection tasks.

III. SOCIAL BEHAVIOR-BASED COMMUNITY DETECTION

In this section, we first formulate the problem of integrating social behaviors, network topology and network content for community detection. Then, we propose a model, namely, social behavior-based community detection (SBCD). Finally, we describe the generative process of the proposed model.

The notations used in our model are shown in Table 1.

TABLE 1. Notations.

Notations	Descriptions
U, K, C, T	set of users, topics, communities and time stamps
W	word set of vocabulary
D_i	posts not on links sent by user i
$E_i, e_{ii'}$	links sent by user i, directed link from user i to i'
W_{ij}, W_{iq}	word list of the j-th post, the q-th link of user i
W_{ijl}, W_{iqr}	the l-th word of word list W_{ij}, the r-th word of word list W_{iq}
$\boldsymbol{\pi}_i$	community distribution (multinomial distribution) of user i
$\boldsymbol{\theta}_c$	topic distribution (multinomial distribution) of community c
$\boldsymbol{\phi}_k$	word distribution (multinomial distribution) of topic k
$\boldsymbol{\psi}_{ik}$	time stamp distribution (multinomial distribution) of user i with topic k
$\boldsymbol{\eta}_{gy,g'y'}$	the probability of forming a link between community g with topic y and community g' with topic y'.
$\boldsymbol{\gamma}_c$	user popularity distribution of community c
c_{ij}, g_{iq}	community indicator of the j-th post and of the q-th link of user i
z_{ij}, y_{iq}	topic indicator of the j-th post and of the q-th link of user i
t_{ij}, t_{iq}	time stamp of the j-th post and of the q-th link of user i
$\alpha, \beta, \epsilon, \rho$	Dirichlet priors

A. PROBLEM FORMULATION

In this section, we state the definitions in our model and summarize the formulation of the problem we solved.

Definition 1: A network is presented by $G = (U, E, D)$. U is the node set. E is the edge set. D is the network content set. $e_{ii'}$ is a directed link from node i to node i'.

Definition 2: Community membership is associated with each user and defined by a vector $\boldsymbol{\pi}$. Its dimension equals the number of communities. For a user i, $\boldsymbol{\pi}_{ic}$ is the probability of belonging to community c. A user in a network may belong to multiple communities, i.e., the user is assigned to all communities with different probabilities. We set a threshold to obtain his real communities. Therefore, our model can detect overlapping communities.

Definition 3: A topic $k \in K$ is defined as a multinomial distribution over vocabularies. For a word w, $\boldsymbol{\phi}_{kw}$ is the probability of belonging to topic k. To investigate community topics, we need to process all documents (e.g., posts in a forum network or papers in a citation network). Given a dataset, we fix its vocabulary with $|W|$ words. A topic is denoted by a set of words. Each word belongs to a topic with a probability $\boldsymbol{\phi}_{kw}$.

Definition 4; Community-topic distribution is defined by a multinomial distribution over topics. For a community c, $\boldsymbol{\theta}_{ck}$ represents the probability of focusing on topic k. A community is a set of users with dense connections. Users in a community might talk about several topics. These topics are used as the semantics of the communities. Therefore, each community focuses on multiple topics.

Definition 5: The time stamp distribution of a user focusing on topic k is defined by a multinomial distribution $\boldsymbol{\psi}_{ik}$ over time stamps. Its dimension is the number of time stamps.

All posts and links are associated with a time stamp in our model. A user's topics of interest might change at different time stamps. Therefore, each topic follows a probability distribution over time stamps.

Definition 6: The user popularity distribution of community c is defined by a multinomial distribution over all users. In community c, $\boldsymbol{\gamma}_{ci}$ represents the probability of user i being interacted by others.

Definition 7: Topic correlation $\boldsymbol{\eta}_{gy,g'y'}$ defines the correlation of topics y and y' in communities g and g', respectively. Suppose that user i is in community g and focuses on topic y, while user i' is in community g' and with topic y'. $\boldsymbol{\eta}_{gy,g'y'}$ means that user i and user i' are likely to communicate with each other when they are in the same community and are interested in the same topic. If g equals g' while y does not equal y', then $\boldsymbol{\eta}_{gy,g'y'}$ is smaller. Although user i and user i' are in the same community, they are interested in different topics. Therefore, the probability of generating a link between them is smaller. For the last two situations (i.e., g does not equal g' and y equals y'; g does not equal g' and y does not equal y'), $\boldsymbol{\eta}_{gy,g'y'}$ should be the smallest value, which means that they are unlikely to communicate with each other if they are not in the same community, regardless of whether they focus on the same topics.

To summarize our model, we formulate the problem we solved as follows: given a graph with rich content, we want to derive each user's community distribution and temporal topic distribution. For communities, we want to identify the topic distribution based on the user's social behaviors, network topology and network content.

B. MODEL STRUCTURE

In this section, we first describe our model structure. Then, we explain its three components in detail to show how the above definitions are implemented to solve the problem we defined.

We propose a generative model to accurately generate network topology, link content and node content by utilizing social behaviors. In this model, (i) user community membership, (ii) the temporal topic distribution of users, (iii) community-topic distribution, and (iv) the word distribution of topics are all latent factors. We want to infer them given network topology and content.

Fig. 2 shows the probabilistic graphical model of SBCD. It consists of three components: a) generation of node content with time stamps; b) generation of link content with time stamps; and c) generation of all links.

1) GENERATION OF NODE CONTENT WITH TIME STAMPS

Considering the second social behavior, i.e., posting preference, we generate only node content in this component. For example, if a user i publishes a tweet that is never replied to/retweeted by others, the tweet is considered as node content. All tweets of user i are generated as follows. Based on user i's community membership distribution $\boldsymbol{\pi}_i$, we sample a community indicator c_{ij} indicating that user i belongs to community c_{ij} when he publishes the j-th tweet. Then, we sample a value of another latent variable based on community-topic distribution $\boldsymbol{\theta}_{c_{ij}}$ denoted by z_{ij}, which means that the topic of the current tweet is z_{ij}. Finally, we generate two observed variables based on topic-word distribution $\boldsymbol{\phi}_{z_{ij}}$ and topic distribution over time $\boldsymbol{\psi}_{iz_{ij}}$: word list and time stamps of the tweet. Here, we utilize the multitopic preference and temporal variation of topics behaviors.

2) GENERATION OF LINK CONTENT WITH TIME STAMPS

Considering the social behavior of posting preference, we generate only link content in this component. Those tweets that are published to reply to others are considered link content. They are generated in the same way as node content generation. For user i's q-th link, we derive its community indicator g_{iq} indicating that user i belongs to community g_{iq} when he sends this post to reply to someone. Then, we sample a value of another latent variable based on community-topic distribution $\boldsymbol{\theta}_{g_{iq}}$ denoted by y_{iq}. This means that the topic of the current link content is y_{iq}. Finally, we generate two observed data based on topic-word distribution $\boldsymbol{\phi}_{y_{iq}}$ and topic distribution over time $\boldsymbol{\psi}_{iy_{iq}}$: word list and time stamps of current link content.

3) GENERATION OF ALL LINKS

For a directed link $e_{ii'}$ that is from i to i', its generation is formulated as follows. $\boldsymbol{\eta}_{gy,g'y'}$ is a factor of generating

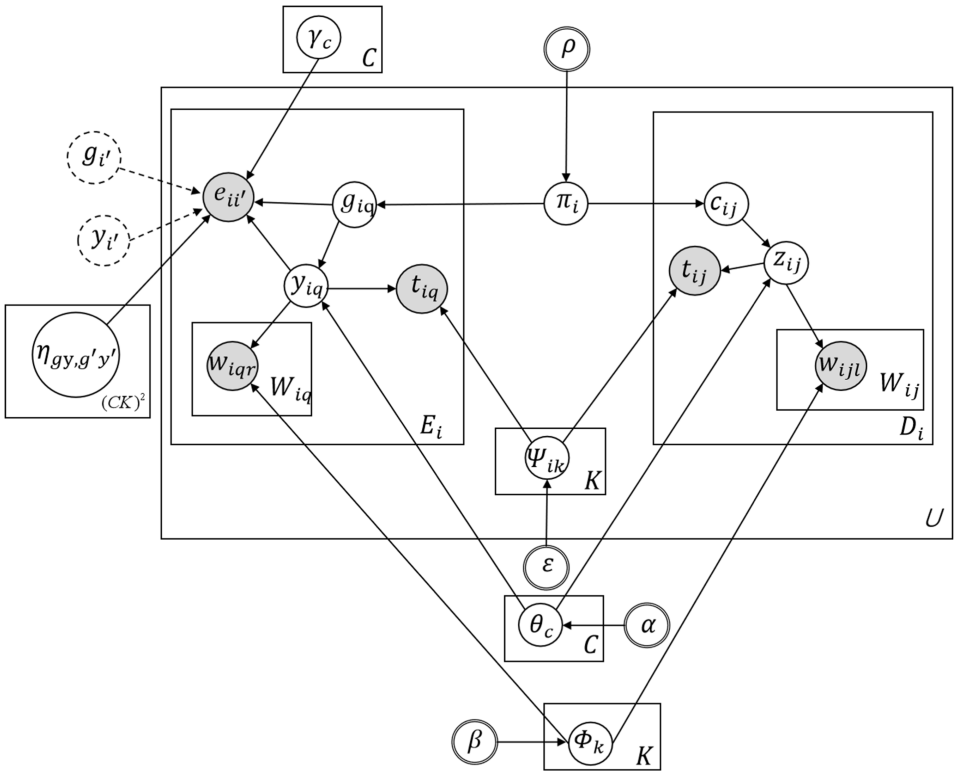

FIGURE 2. The graphical representation of our model.

$e_{ii'}$ because it represents topic correlation with respect to communities. Moreover, the social behavior of reciprocity of interactions indicates that users tend to focus on popular or authoritative users. To integrate the above two factors, we define $\omega_{ii'}$ as follows:

$$\omega_{ii'} = \boldsymbol{\eta}_{gy,g'y'} + \boldsymbol{\gamma}_{y'i'} + \lambda_i, \tag{1}$$

where y' is the community indicator of user i' and λ_i denotes the characteristic (i.e., active or inactive) of user i. λ_i is calculated by $\frac{(out-degree)_i}{(degree)_i}$.

Then, a sigmoid function is adopted to calculate the probability of this link.

$$\begin{aligned} P(E^t_{ii'} &= 1|g, g', y, y', \boldsymbol{\eta}, \boldsymbol{\gamma}) \\ &= \sigma(\omega_{ii'}) \\ &= 1/(1 + e^{-\omega_{ii'}}) \end{aligned} \tag{2}$$

Because we use the Gibbs sampling method for inference, which makes it difficult to process the sigmoid function, we adopt the P'olya-Gamma distribution to model the sigmoid function [58].

$$\frac{1}{1 + e^{-\omega_{ij}}} = \frac{1}{2} \int_0^\infty \varphi(\omega_{ij}, \xi_{ij}) P(\xi_{ij}) d\xi_{ij} \tag{3}$$

where $\varphi(\omega_{ij}, \xi_{ij}) = e^{(\omega_{ij} - \xi_{ij}\omega_{ij}^2)/2}$ and $\xi_{ij} \sim PG(1, 0)$. Then, we derive a joint probability distribution:

$$P(E^t_{ii'} = 1, \xi_{ii'}) = \frac{1}{2} \varphi(\omega_{ii'}, \xi_{ii'}) P(\xi_{ii'}|1, 0) \tag{4}$$

C. GENERATIVE PROCESS

Summarizing the above components, the generative process of SBCD is described as follows.

1) Initialize $\boldsymbol{\eta}$ randomly;
2) For each topic $k \in K$,
   a) Sample word distribution from a Dirichlet prior: $\boldsymbol{\phi}_k \mid \beta \sim Dir(\beta)$;
3) For each community $c \in C$,
   a) Sample distribution over topics from a Dirichlet prior: $\boldsymbol{\theta}_c \mid \alpha \sim Dir(\alpha)$;
   b) Initialize γ_c with a Uniform distribution;
4) For each user $i \in U$,
   a) Sample his community distribution from a Dirichlet prior: $\boldsymbol{\pi}_i \mid \rho \sim Dir(\rho)$;
   b) For each topic $k \in K$,
      i) Sample distribution over time stamps from a Dirchlet prior: $\boldsymbol{\psi}_{ik} \mid \varepsilon \sim Dir(\varepsilon)$
5) For each user $i \in U$,
   a) For each post $j \in D_i$,
      i) Sample community indicator from a Multinomial distribution: $c_{ij} \mid \boldsymbol{\pi}_i \sim Mul(\boldsymbol{\pi}_i)$;
      ii) Sample topic indicator from a Multinomial distribution: $z_{ij} \mid \boldsymbol{\theta}_{c_{ij}} \sim Mul(\boldsymbol{\theta}_{c_{ij}})$;
      iii) For each word $l \in W_{ij}$,
      - Sample word from a Multinomial distribution: $w_{ijl} \mid \boldsymbol{\phi}_{z_{ij}} \sim Mul(\boldsymbol{\phi}_{z_{ij}})$;
      iv) Sample time stamp $t_{ij} \mid \boldsymbol{\psi}_{iz_{ij}} \sim Mul(\boldsymbol{\psi}_{iz_{ij}})$;

b) For each link $q \in E_i$,
   i) Sample community indicator from a Multinomial distribution: $g_{iq} \mid \boldsymbol{\pi}_i \sim Mul(\boldsymbol{\pi}_i)$;
   ii) Sample topic indicator from a Multinomial distribution: $y_{iq} \mid \boldsymbol{\theta}_{g_{iq}} \sim Mul(\boldsymbol{\theta}_{g_{iq}})$;
   iii) Sample the link from i to i': $E^t_{ii'} \mid g_{iq}, g_{i'}, y_{iq}, y_{i'}, \boldsymbol{\eta}, \boldsymbol{\gamma} \sim Ber(\sigma(\boldsymbol{\eta}_{g_{iq}y_{iq},g_{i'}y_{i'}} + \boldsymbol{\gamma}_{i'g_{i'}})$;
   iv) For each word $r \in W_{iq}$,
   - Sample word from a Multinomial distribution: $w_{iqr} \mid \boldsymbol{\phi}_{y_{iq}} \sim Mul(\boldsymbol{\phi}_{y_{iq}})$;

   v) Sample time stamp $t_{iq} \mid \boldsymbol{\psi}_{iy_{iq}} \sim Mul(\boldsymbol{\psi}_{iy_{iq}})$;

IV. MODEL INFERENCE

In this model, U, E, and D are observed data. Our target is to estimate $\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{\gamma}$. The full posterior distribution of SBCD is:

$$\begin{aligned} &P(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{\gamma}, c, z, g, y, \xi | U, E, D, \rho, \alpha, \beta, \varepsilon, t) \\ &\quad \propto P(\boldsymbol{\pi}|\rho)P(\boldsymbol{\theta}|\alpha)P(\boldsymbol{\phi}|\beta)P(\boldsymbol{\psi}|\varepsilon)P(c, g|\boldsymbol{\pi}) \\ &\qquad \cdot P(z|c, \boldsymbol{\theta})P(w_d|z, \boldsymbol{\phi})P(t_d|z, \boldsymbol{\psi}) \\ &\qquad \cdot P(y|g, \boldsymbol{\theta})P(w_e|y, \boldsymbol{\phi})P(t_e|y, \boldsymbol{\psi}) \cdot P(e, \xi|\boldsymbol{\gamma}, \boldsymbol{\eta}, g, y). \end{aligned} \tag{5}$$

Because it is difficult to calculate the normalizing constant, we adopt collapsed Gibbs sampling [59] for approximate inference.

A. APPROXIMATE INFERENCE

First, we marginalize out all parameters, i.e., $\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}$, and $\boldsymbol{\psi}$.

$$\begin{aligned} P(c, z, g, y|.) \propto &\int P(\boldsymbol{\pi}|\rho)P(c, g|\boldsymbol{\pi})d\boldsymbol{\pi} \\ &\cdot \int P(\boldsymbol{\theta}|\alpha)P(z|c, \boldsymbol{\theta})P(y|g, \boldsymbol{\theta})d\boldsymbol{\theta} \\ &\cdot \int P(\boldsymbol{\phi}|\beta)P(w_d|z, \boldsymbol{\phi})P(w_e|y, \boldsymbol{\phi})d\boldsymbol{\phi} \\ &\cdot \int P(\boldsymbol{\psi}|\varepsilon)P(t_d|z, \boldsymbol{\psi})P(t_e|y, \boldsymbol{\psi})d\boldsymbol{\psi} \\ &\cdot P(e, \xi)). \end{aligned} \tag{6}$$

Second, we calculate all integrals in (6). The first integral is calculated by (7).

$$\begin{aligned} &\int P(\boldsymbol{\pi}|\rho)P(c, g|\boldsymbol{\pi})d\boldsymbol{\pi} \\ &= \int (\prod_{i=1}^{|U|} \frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{|C|}} \prod_{c=1}^{|C|} \pi_{ic}^{\rho-1})(\prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{c=1}^{|C|} \pi_{ic}^{n_j^{(c)}}) \\ &\quad \cdot (\prod_{i=1}^{|U|}\prod_{q=1}^{|E_i|}\prod_{c=1}^{|C|} \pi_{ig}^{n_q^{(g)}})d\pi \\ &= \prod_{i=1}^{|U|} \frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^C} \int \prod_{i=1}^{|U|}\prod_{c=1}^{|C|} \pi_{ic}^{n_i^{(c)}+\rho-1} d\pi \\ &= \prod_{i=1}^{|U|} \frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{|C|}} \cdot \frac{\prod_{c=1}^{|C|} \Gamma(n_i^{(c)} + \rho)}{\Gamma(n_i^{(\cdot)} + |C|\rho)}, \end{aligned} \tag{7}$$

where n_i^c is the number of posts and links that are assigned to community c for user i. $n_i^{\cdot}$ is the number of posts and links that are assigned to all communities.

The second integral is calculated as follows.

$$\begin{aligned} &\int P(\boldsymbol{\theta}|\alpha)P(z|c, \boldsymbol{\theta})P(y|g, \boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int (\prod_{c=1}^{|C|} \frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^K} \prod_{k=1}^{|K|} \theta_{ck}^{\alpha-1}) \\ &\quad \cdot \prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{k=1}^{|K|} \theta_{ck}^{n_{jc}^{(k)}} \prod_{i=1}^{|U|}\prod_{q=1}^{|E_i|}\prod_{k=1}^{|K|} \theta_{gk}^{n_{qg}^{(k)}} d\theta \\ &= \prod_{c=1}^{|C|} \frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^K} \int \prod_{c=1}^{|C|}\prod_{k=1}^{|K|} \theta_{ck}^{n_{Dc}^{(k)}+n_{Ec}^{(k)}+\alpha-1} d\theta \\ &= \prod_{c=1}^{|C|} \frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^{|K|}} \cdot \frac{\prod_{k=1}^{|K|} \Gamma(n_{Dc}^{(k)} + n_{Ec}^{(k)} + \alpha)}{\Gamma(n_{Dc}^{(\cdot)} + n_{Ec}^{(\cdot)} + |K|\alpha)}, \end{aligned} \tag{8}$$

where $n_{Dc}^{(k)}$ is the number of all posts that are assigned to community c with topic k. $n_{Ec}^{(k)}$ corresponds to all links. $n_{Dc}^{(\cdot)}$ and $n_{Ec}^{(\cdot)}$ integrate all topics. The third integral is calculated as follows.

$$\begin{aligned} &\int P(\boldsymbol{\phi}|\beta)P(w_d|z, \boldsymbol{\phi})P(w_e|y, \boldsymbol{\phi})d\boldsymbol{\phi} \\ &= \int (\prod_{k=1}^{|K|} \frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}} \prod_{w=1}^{|W|} \phi_{kw}^{\beta-1}) \prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{w=1}^{|W|} \phi_{jz}^{n_{zw}^{(w)}} \\ &\quad \cdot \prod_{i=1}^{|U|}\prod_{q=1}^{|E_i|}\prod_{w=1}^{|W|} \phi_{yw}^{n_{qy}^{(w)}} d\phi \\ &= \prod_{k=1}^{|K|} \frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}} \int \prod_{k=1}^{|K|}\prod_{w=1}^{|W|} \phi_{kw}^{n_{Dk}^{(w)}+n_{Ek}^{(w)}+\beta-1} d\phi \\ &= \prod_{k=1}^{|K|} \frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}} \cdot \frac{\prod_{w=1}^{|W|} \Gamma(n_{Dk}^{(w)} + n_{Ek}^{(w)} + \beta)}{\Gamma(n_{Dk}^{(\cdot)} + n_{Ek}^{(\cdot)} + |W|\beta)}, \end{aligned} \tag{9}$$

where $n_{Dk}^{(w)}$ is the number of times assigned to topic k for word w in all posts. $n_{Ek}^{(w)}$ corresponds to all links. $n_{Dk}^{(\cdot)}$ and $n_{Ek}^{(\cdot)}$ integrate all words. For the last integral in (6),

$$\begin{aligned} &\int P(\boldsymbol{\psi}|\varepsilon)P(t_d|z, \boldsymbol{\psi})P(t_e|y, \boldsymbol{\psi})d\boldsymbol{\psi} \\ &= \int \prod_{k=1}^{|K|} \frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}} \prod_{t=1}^{|T|} \psi_{ck}^{\varepsilon-1} \prod_{i=1}^{|U|}\prod_{j=1}^{|D_i|}\prod_{t=1}^{|T|} \psi_{cz}^{n_{jcz}^{(t)}} \\ &\quad \cdot \prod_{i=1}^{|U|}\prod_{q=1}^{|E_i|}\prod_{t=1}^{|T|} \psi_{gy}^{n_{qgy}^{(t)}} d\psi \\ &= \prod_{k=1}^{|K|} \frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}} \cdot \int \prod_{k=1}^{|K|}\prod_{t=1}^{|T|} \psi_{ck}^{n_{Dck}^{(t)}+n_{Eck}^{(t)}+\varepsilon-1} d\psi \\ &= \prod_{k=1}^{|K|} \frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}} \cdot \frac{\prod_{t=1}^{T} \Gamma(n_{Dck}^{(t)} + n_{Eck}^{(t)} + \varepsilon)}{\Gamma(n_{Dck}^{(\cdot)} + n_{Eck}^{(\cdot)} + |T|\varepsilon)}, \end{aligned} \tag{10}$$

where $n_{Dck}^{(t)}$ is the number of posts that are assigned to community c with topic k at time stamp t. $n_{Eck}^{(t)}$ corresponds to all links. $n_{Dck}^{(\cdot)}$ and $n_{Eck}^{(\cdot)}$ integrate all time stamps.

Third, we sample all latent variables, i.e., c, z, g, y, and ξ. For user i's post d_{ij}, its community membership c_{ij} and topic indicator k_{ij} are sampled as follows.

$$\begin{aligned}
&P(c_{ij}c|c_{\neg ij}, z_{ij}=k, t_{ij}=t, g, y, .)\\
&\quad=\frac{P(c,z,g,y)}{P(c_{\neg ij},z,g,y)}\\
&\quad=\frac{\int P(\boldsymbol{\pi}|\rho)P(c,g|\boldsymbol{\pi})d\boldsymbol{\pi}}{\int P(\boldsymbol{\pi}|\rho)P(c_{\neg ij},g|\boldsymbol{\pi})d\boldsymbol{\pi}}\\
&\qquad\cdot\frac{\int P(\boldsymbol{\theta}|\alpha)P(z|c,\boldsymbol{\theta})P(y|g,\boldsymbol{\theta})d\boldsymbol{\theta}}{\int P(\boldsymbol{\theta}|\alpha)P(z|c_{\neg ij},\boldsymbol{\theta})P(y|g,\boldsymbol{\theta})d\boldsymbol{\theta}}\\
&\quad=\frac{n^{(c)}_{i,\neg ij}+\rho}{n^{(\cdot)}_{i,\neg ij}+|C|\rho}\cdot\frac{n^{(k)}_{c,\neg ij}+\alpha}{n^{(\cdot)}_{c,\neg ij}+|K|\alpha},
\end{aligned}\tag{11}$$

where $n^{(c)}_{i,-ij}$ is the number of posts and links that are assigned to community c excluding post d_{ij}. $n^{(k)}_{c,-ij}$ is the number of posts and links that are assigned to community c with topic k excluding post d_{ij}. $n^{(t)}_{ck,-ij}$ is the number of posts and links with community c and topic k that appear at time stamp t excluding post d_{ij}. Dots denote marginal counts.

$$\begin{aligned}
&P(z_{ij}=k|z_{\neg ij}, c_{ij}=c, t_{ij}=t, g, y, .)\\
&\quad=\frac{P(z,c,g,y)}{P(z_{\neg ij},c,g,y)}\\
&\quad=\frac{\int P(\boldsymbol{\theta}|\alpha)P(z|c,\boldsymbol{\theta})P(y|g,\boldsymbol{\theta})d\boldsymbol{\theta}}{\int P(\boldsymbol{\theta}|\alpha)P(z_{\neg ij}|c,\boldsymbol{\theta})P(y|g,\boldsymbol{\theta})d\boldsymbol{\theta}}\\
&\qquad\cdot\frac{\int P(\boldsymbol{\phi}|\beta)P(w_d|z,\boldsymbol{\phi})P(w_e|y,\boldsymbol{\phi})d\boldsymbol{\phi}}{\int P(\boldsymbol{\phi}|\beta)P(w_d|z_{\neg ij},\boldsymbol{\phi})P(w_e|y,\boldsymbol{\phi})d\boldsymbol{\phi}}\\
&\qquad\cdot\frac{\int P(\boldsymbol{\psi}|\varepsilon)P(t_d|z,\boldsymbol{\psi})P(t_e|y,\boldsymbol{\psi})d\boldsymbol{\psi}}{\int P(\boldsymbol{\psi}|\varepsilon)P(t_d|z_{\neg ij},\boldsymbol{\psi})P(t_e|y,\boldsymbol{\psi})d\boldsymbol{\psi}}\\
&\quad=\frac{n^{(k)}_{c,\neg ij}+\alpha}{n^{(\cdot)}_{c,\neg ij}+|K|\alpha}\\
&\qquad\cdot\frac{\Pi_{w=1}^{|W|}\Pi_{q=0}^{n^{(w)}_{ij}-1}(n^{(w)}_{k,\neg ij}+q+\beta)}{\Pi_{q=0}^{n^{(\cdot)}_{ij}-1}(n^{(\cdot)}_{k,\neg ij}+q+\beta)}\cdot\frac{n^{(t)}_{ck,\neg ij}+\varepsilon}{n^{(\cdot)}_{ck,\neg ij}+|T|\varepsilon},
\end{aligned}\tag{12}$$

For each link $e_{ii'}$, suppose that it is the q-th link of user i. User i's community membership g_{iq} is sampled by (13).

$$\begin{aligned}
&P(g_{iq}=c|g_{\neg iq}, y_{iq}=k, t_{iq}=t, c, z, .)\\
&\quad=\frac{P(g,c,z,y)}{P(g_{\neg iq},c,z,y)}\\
&\quad=\frac{\int P(\boldsymbol{\pi}|\rho)P(c,g|\boldsymbol{\pi})d\boldsymbol{\pi}}{\int P(\boldsymbol{\pi}|\rho)P(c,g_{\neg iq}|\boldsymbol{\pi})d\boldsymbol{\pi}}\\
&\qquad\cdot\frac{\int P(\boldsymbol{\theta}|\alpha)P(z|c,\boldsymbol{\theta})P(y|g,\boldsymbol{\theta})d\boldsymbol{\theta}}{\int P(\boldsymbol{\theta}|\alpha)P(z|c,\boldsymbol{\theta})P(y|g_{\neg iq},\boldsymbol{\theta})d\boldsymbol{\theta}}\\
&\qquad\cdot\varphi(\omega_{iq},\xi_{iq})\\
&\quad=\frac{n^{(c)}_{i,\neg iq}+\rho}{n^{(\cdot)}_{i,\neg iq}+|C|\rho}\cdot\frac{n^{(k)}_{c,\neg iq}+\alpha}{n^{(\cdot)}_{c,\neg iq}+|K|\alpha}\\
&\qquad\cdot\varphi(\omega_{iq},\xi_{iq}).
\end{aligned}\tag{13}$$

Each link's topic y_{iq} is sampled as follows.

$$\begin{aligned}
&P(y_{iq}=k|y_{\neg iq}, g_{iq}=c, t_{iq}=t, c, z, .)\\
&\quad=\frac{P(y,c,z,g)}{P(y_{\neg iq},c,z,g)}\\
&\quad=\frac{\int P(\boldsymbol{\theta}|\alpha)P(z|c,\boldsymbol{\theta})P(y|g,\boldsymbol{\theta})d\boldsymbol{\theta}}{\int P(\boldsymbol{\theta}|\alpha)P(z|c,\boldsymbol{\theta})P(y_{\neg iq}|g,\boldsymbol{\theta})d\boldsymbol{\theta}}\\
&\qquad\cdot\frac{\int P(\boldsymbol{\phi}|\beta)P(w_d|z,\boldsymbol{\phi})P(w_e|y,\boldsymbol{\phi})d\boldsymbol{\phi}}{\int P(\boldsymbol{\phi}|\beta)P(w_d|z,\boldsymbol{\phi})P(w_e|y_{\neg iq},\boldsymbol{\phi})d\boldsymbol{\phi}}\\
&\qquad\cdot\frac{\int P(\boldsymbol{\psi}|\varepsilon)P(t_d|z,\boldsymbol{\psi})P(t_e|y,\boldsymbol{\psi})d\boldsymbol{\psi}}{\int P(\boldsymbol{\psi}|\varepsilon)P(t_d|z,\boldsymbol{\psi})P(t_e|y_{\neg iq},\boldsymbol{\psi})d\boldsymbol{\psi}}\\
&\qquad\cdot\varphi(\omega_{iq},\xi_{iq})\\
&\quad=\frac{n^{(k)}_{c,\neg iq}+\alpha}{n^{(\cdot)}_{c,\neg iq}+|K|\alpha}\\
&\qquad\cdot\frac{\Pi_{w=1}^{|W|}\Pi_{v=0}^{n^{(w)}_{iq}-1}(n^{(w)}_{k,\neg iq}+v+\beta)}{\Pi_{v=0}^{n^{(\cdot)}_{iq}-1}(n^{(\cdot)}_{k,\neg iq}+v+\beta)}\\
&\qquad\cdot\frac{n^{(t)}_{ck,\neg iq}+\varepsilon}{n^{(\cdot)}_{ck,\neg iq}+|T|\varepsilon}\cdot\varphi(\omega_{iq},\xi_{iq}).
\end{aligned}\tag{14}$$

Finally, we sample $\xi_{ii'}$.

$$P(\xi_{ii'}|.)\propto e^{-\frac{1}{2}\xi_{ii'}\omega^2_{ii'}}P(\xi_{ii'}|1,0)=PG(1,\omega_{ii'}).\tag{15}$$

B. PARAMETER ESTIMATION

After the Gibbs sampler converges, all parameters are estimated as follows.

$$\hat{\boldsymbol{\pi}}_{ic}=\frac{n^{(c)}_{i}+\rho}{n^{(\cdot)}_{i}+|C|\rho}\tag{16}$$

$$\hat{\boldsymbol{\theta}}_{ck}=\frac{n^{(k)}_{c}+\alpha}{n^{(\cdot)}_{c}+|K|\alpha}\tag{17}$$

$$\hat{\boldsymbol{\phi}}_{kw}=\frac{n^{(w)}_{k}+\beta}{n^{(\cdot)}_{k}+|W|\beta}\tag{18}$$

$$\hat{\boldsymbol{\psi}}_{kc,t}=\frac{n^{(t)}_{ck}+\varepsilon}{n^{(\cdot)}_{ck}+|T|\varepsilon}\tag{19}$$

Parameter $\boldsymbol{\eta}$ is calculated by aggregating all community-topic pairs with respect to all links. Parameter $\boldsymbol{\gamma}$ for each community is calculated by counting the number of links whose target node is in the current community.

C. ALGORITHM SUMMARIZATION AND TIME COMPLEXITY

The inference procedure is shown in the following algorithm. Next, we analyze the time complexity of the algorithm of SBCD. As shown later, it runs linearly in terms of network data (i.e., the number of users, links, topics, vocabulary).

T denotes the number of iterations for convergence. All counters (e.g., how many times a user is assigned to a community) are recorded in memory. In Steps 3-7, community

Algorithm 1 Inference for SBCD

Require: users U, user posts D with time stamps, links E with content and time stamps;

Ensure: community distribution $\boldsymbol{\pi}$, topic distribution $\boldsymbol{\theta}$, word distribution $\boldsymbol{\phi}$, topic distribution over time $\boldsymbol{\psi}$, parameter $\boldsymbol{\eta}$, parameter $\boldsymbol{\gamma}$;

```
Initialize α, β, ε, ρ, η, γ;
for iter = 1 : T do
    for each user i ∈ U do
        for each post d_ij ∈ D_i do
            Sample community indicator c_ij according to (11);
            Sample topic indicator z_ij according to (12);
        end for
        for each link e_ii' ∈ E_i do
            Sample community indicator g_iq according to (13);
            Sample topic indicator y_iq according to (14);
            Sample ξ_ii' according to (15);
        end for
    end for
    for each link e ∈ E do
        Update η and γ by aggregating community and topic of two endpoint users;
    end for
end for
Caculate π̂, θ̂, φ̂, and ψ̂ according to (16) - (19);
```

indicators and topic indicators for posts of all users are sampled. Steps 5 takes constant time. In Step 6, it takes $\Theta(|W|)$ to compute the second fraction of (12) for a specific topic, where $|W|$ is the vocabulary size. Thus, Steps 3-7 take $\Theta(|U|\times|D|\times|C|+|U|\times|D|\times|K|\times|W|)$. Steps 8-12 compute the community indicator, topic indicator and ξ. The number of all links is $|E|$. Equations (13), (14) and (15) take constant time. Thus, Steps 8-12 take $\Theta(|E|\times|C|+|E|\times|K|\times|W|)$. For Steps 14-16, we calculate $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$. It takes $\Theta(|E|)$. Based on the above discussions, the complexity of SBCD is linear to the data size. As datasets become large, we can parallelize our model.

V. EXPERIMENTS

In this section, we evaluate our model's accuracy for community detection and show a case study to illustrate the social behaviors that are investigated. We choose two real datasets and six state-of-the-art baselines. Our experiments are conducted on a personal computer with an Intel Core i7-7700K @ 4.2 GHz CPU and 64 GB RAM.

A. DATASETS

To evaluate the accuracy of the community detection results, we choose two real networks that include all four types of social behaviors. One of them is the Reddit dataset, and the other is the DBLP dataset [60].

1) REDDIT DATASET

Posts are crawled from three sub-forums of reddit.com (i.e., *Science*, *movie*, and *Politics*). These three sub-forums correspond to three communities. The author of each post (including reply comment) is extracted as a node. The sub-forum to which the post belongs is selected as ground-truth community of the author. Therefore, the ground-truth only depends on the truth where a node really appears. Moreover, when a node exists in multiple sub-forums he belongs to multi communities and the communities are overlapped. When a user replies to a post of others, a directed link is generated. Main threads are considered as node content, and posts that are replies to other posts are used as link content. All posts are recorded with time stamps. We divide the dataset into seven snapshots with one day as a time window.

2) DBLP DATASET

It is a paper coauthorship network. The papers are crawled from three research fields, i.e., *Machine Learning*, *Image Processing*, and *Data Mining* corresponding to three communities. The authors are considered as nodes. When several authors publish one paper corporately, links are created among them. The research topics are considered as ground-truth community of the authors. When an author has multiple research fields the communities are overlapped. Therefore, the ground-truth reveals the truth of authors' research fields. When several authors publish one paper corporately, links are created among them. The title of a paper is used as both node content and link content. The dataset is divided into eleven snapshots with one year as a time window.

All datasets are processed by removing stop words and stemming. They are summarized in Table 2.

B. BASELINES

Our model integrates network topology, network content, and user social behaviors for community detection. Therefore, we choose six similar state-of-the-art baselines to evaluate our model's accuracy. Some of them model some user attributes (e.g., documents, roles) to detect the community structure of networks. They are all generative models. The six baselines are described as follows:

- **Community Level Diffusion (COLD)** [48]. It is a generative model that generates network topology and content based on the latent community membership factor. It can identify the diffusion pattern between communities.
- **Community Profiling and Detection (CPD)** [49]. It integrates friendship relations, diffusion links and individual preferences to identify community profiling. This is the first work to propose community profiling.
- **Poisson Mixed-Topic Link Model (PMTLM)** [61]. It combines the LDA model and Poisson distribution to generate network topology and content. It generates documents and links based on the same latent factor. Because it detects the community structure of

TABLE 2. Summarization of datasets.

	#users	#links	#words	#snapshots	#communities	#topics
Reddit	38,602	96,476	13,223	7	3	3
DBLP	20,322	832,611	8,123	11	3	3

TABLE 3. Experimental results comparations on Reddit and DBLP.

Metrics (%)	Datasets	Methods						
		COLD	CPD	CRM	PMTLM	GHIPT	TCCD	Ours
GNMI	Reddit	16.24	13.12	38.47	41.61	60.81	59.07	**61.98**
	DBLP	31.77	25.56	20.99	14.94	45.13	44.23	**46.35**
F-score	Reddit	59.81	80.49	69.96	64.30	83.53	82.95	**84.17**
	DBLP	74.90	78.66	70.50	72.01	82.26	81.50	**83.65**
Jaccard	Reddit	44.82	70.68	56.49	57.36	73.16	72.64	**73.53**
	DBLP	60.45	65.26	55.39	56.28	71.35	69.10	**72.37**

a document network, we need to integrate the documents of each user to infer his or her community membership.

- **Community Role Model (CRM)** [52]. It assigns roles to users. Friendship links and diffusion links are modeled in networks based on users' community assignment.
- **Community Detection Considering Group Homophily and Individual Personality of Topics (GHIPT)** [62]. It proposes a novel generative community detection model by integrating group homophily and individual personality of topics.
- **Topic Correlations-Based Community Detection (TCCD)** [63]. TCCD is proposed by our previous work that is extended. It considers the correlations of different topics in the community detection model.

C. METRICS

The output of our method for each node is the probability distribution over communities. We set a threshold (e.g., 0.33 for three communities) to get the community label of nodes. Therefore, we can get overlapping community structure. Since both datasets supply ground truth, we use generalized normalized mutual information (GNMI), F score and the Jaccard index as metrics to evaluate the accuracy of the community detection results. For a dataset with ground truth, the GNMI measure is used to evaluate the accuracy of overlapping community structures [64]. The F score is the harmonic mean of precision and recall: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. The Jaccard index is used to measure the similarity of sample sets A and B, i.e., $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

D. PARAMETER INITIATION

The community number and topic number are set to true values according to the ground truth. η is initiated with random values. Existing works have proven that Dirichlet hyper-parameters have low impact on the efficiency of generative models of our work. Moreover, empirical studies also show that our model is insensitive to hyper-parameters. Therefore, all Dirichlet hyperparameters are initiated by fixed values according to the common strategy (i.e., $\rho = 0.01$, $\alpha = 0.001$, $\beta = 0.1$, $\varepsilon = 0.001$) [48], [65], [66]. We set the threshold for determining community memberships to $1/|C|$. All parameters of the baselines are set to recommended values by the authors.

E. COMPARISON WITH BASELINES

Table 3 shows the comparisons between SBCD and the baselines on two datasets. Overall, SBCD outperforms the other baselines for all metrics.

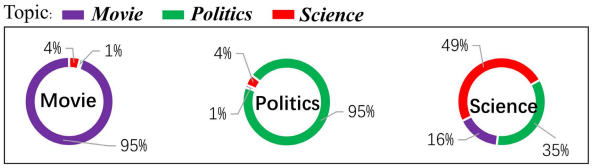

FIGURE 3. Topic distribution of communities on Reddit. Doughnut charts represent communities and colors denote topics.

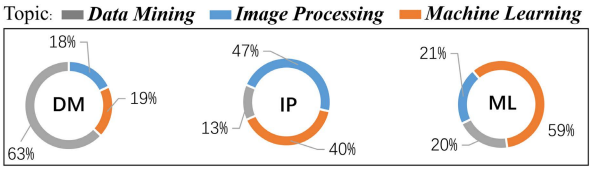

FIGURE 4. Topic distribution of communities on DBLP. *Data Mining*, *Image Processing*, and *Machine Learning* are presented by "DM", "IP", and "ML" for short.

On the Reddit dataset, there are 3,925 isolated posts, i.e., node content. They are not on links. Therefore, they do not contribute to the generation of links. Our model separates these isolated posts from link posts, such that they do not participate in the formation of network topology. COLD uses all posts to generate links, and it does not discriminate link content and node content. The PMTLM must eliminate isolated posts, or there will be an error. CPD does not consider link content at all, which means that all posts are used as node content and provide no useful information to accurately generate network topology. The results show that SBCD achieves 1.92%, 0.77%, and 0.51% improvements in terms of GNMI, F score and Jaccard over the second baseline, i.e., GHIPT. Compared with our previous work, TCCD published in AAAI, the results are improved by 4.93%, 1.47%, and 1.23% in terms of GNMI, F score, and Jaccard, respectively.

On the DBLP dataset, it is quite different from the situation on Reddit. The title of a paper is used by all its authors as node content. Meanwhile, it is also used as the content of the link from one author to another author. Therefore, except for the situation when a paper has only one author, which is a

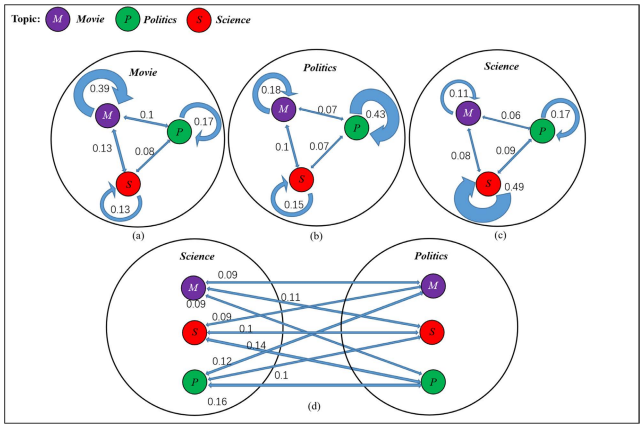

FIGURE 5. Topic correlations on Reddit. Big circles denote communities. Small solid circles with colors denote topics. Arrows present topic correlations inside a community and between communities. (a) Topic correlations in *Movie*. (b) Topic correlations in community *Politics*. (c) Topic correlations in community *Science*. (d) Topic correlations between communities *Science* and *Politics*.

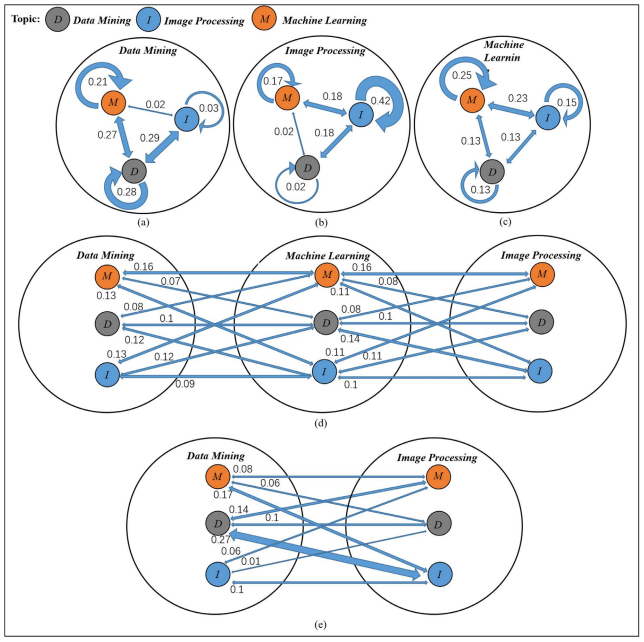

FIGURE 6. Topic correlations on DBLP. (a) Topic correlations in *Data Mining*. (b) Topic correlations in *Image Processing*. (c) Topic correlations in *Machine Learning*. (d) Topic correlations between communities *Data Mining* and *Machine Learning*; topic correlations between communities *Image Processing* and *Machine Learning*. (e) Topic correlations between communities *Data Mining* and *Image Processing*.

very uncommon case, node content is used as link content. The results show that SBCD achieves 4.79%, 2.64%, and 4.73% improvements in terms of GNMI, F score and Jaccard, respectively, over TCCD. It achieves a 1.92% improvement in terms of GNMI over the second-best baseline, i.e., GHIPT. The experimental results show that considering social behaviors in SBCD enables better results to be achieved. The results are summarized as follows.

1) Although all baselines utilize topology and network content, our model derives a more accurate community structure by considering underlying factors, i.e., social behaviors that lead to the generation of communities. The results show that the social behaviors we proposed have profound impacts on community structure.

2) Moreover, for those networks without much isolated node content, such as paper coauthorship networks, SBCD outperforms all baselines for all metrics.

F. CASE STUDY

In addition to community structure, we are also interested in community semantics, i.e., the topic distribution of communities, topic correlations between communities and word distribution of topics. Moreover, by considering social behaviors, we further illustrate the following significant information of users on both datasets: user popularity and topic changes of users.

1) TOPIC DISTRIBUTION OF COMMUNITIES

As Fig. 3 shows, the topics *movie* and *Politics* are dominant in the communities *movie* and *Politics*, respectively. However, for the community *Science*, although the topic *Science* is dominant, there are 35 percent of posts talking about *Politics* and 16 percent of posts talking about *movie*. Fig. 4 shows that the topics *Data Mining*, *Image Processing*, and *Machine Learning* are dominant in the communities *Data Mining*, *Image Processing*, and *Machine Learning*, respectively. Compared with the Reddit dataset, it is true that a research field involves multiple topics, as studies become increasingly cross-disciplinary.

2) TOPIC CORRELATIONS

Fig. 5(a), Fig. 5(b), and Fig. 5(c) illustrate topic correlations inside the communities *movie*, *Politics*, and *Science*, respectively. As they show, all topics are discussed in the three communities with different weights. The above subsection shows that the topics *movie*, *Politics*, and *Science* are dominant inside the corresponding communities. Beyond the above observations, Fig. 5 reveals significant information that users with different topic interests also communicate with each other with different probabilities. The interactions between users who are interested in *movie*, *Politics*, and *Science* account for the largest proportion, i.e., 0.39, 0.43, and 0.49, respectively. Users focusing on other topics also interact with each other but with less intense communication. Fig. 5(d) shows topic correlations between the communities *Science* and *Politics*. Users focusing on the topic *Politics* in the community *Science* have a larger probability of communicating with users who also focus on the topic *Politics* in the community *Politics*. The second-largest probability of communication is between users focusing on the topic *Science* in the community *Science* and users focusing on the topic *Politics* in the community *Politics*. The topic correlations between other community pairs are not presented because the probabilities of communication are all smaller than 0.01. This means that users in other community pairs seldom interact with each other.

Fig. 6 shows topic correlations on the DBLP dataset. Fig. 6(a), Fig. 6(b), and Fig. 6(c) illustrate topic correlations inside the communities of *Data Mining*, *Image Processing*, and *Machine Learning*, respectively. The topics *Data Mining*, *Image Processing*, and *Machine Learning* are dominant topics in the corresponding communities. Moreover, *Machine Learning* is a hot research topic in all communities. Fig. 6(d) and Fig. 6(e) show the topic correlations across the three communities. They reveal a true phenomenon, that authors in the *Image Processing* and *Data Mining* research fields have intense communications with authors in the *Machine Learning* field.

3) WORD DISTRIBUTION OF TOPICS

We use word clouds to illustrate topics. As shown in Fig. 7 and Fig. 8, all topics identified by SBCD are meaningful.

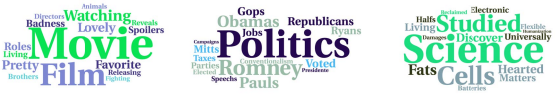

FIGURE 7. Word clouds of three topics: ***Movie***, ***Politics*** and ***Science***.

FIGURE 8. Word clouds of three topics: ***Data Mining***, ***Image Processing***, and ***Machine Learning***.

TABLE 4. Top 5 authors in each community on DBLP dataset.

Communities	Top 5 users
Data mining	Philip S. Yu; Yufei Tao; Nick Koudas; Surajit Chaudhuri; Jiawei Han
Image processing	Thomas S. Huang; Xiaoou Tang; William T. Freeman; Trevor Darrell; Shuicheng Yan;
Machine learning	Toby Walsh; Michael I. Jordan; Vincent Conitzer; Thomas Seidl; Zoubin Ghahramani

4) USER POPULARITY

As described in the first social behavior, users with a large influence often attract more attention. This case study verifies "user popularity" social behavior by analyzing parameter γ which denotes the popularity of users in each community and is responsible for the generation of links. If a node is popular or authoritative, there will be a large amount of links pointing to him. Therefore, it is the integration of a node's popularity and the contents he published affect community structure and community semantics.

Top 5 authors are selected in each community on the DBLP dataset, as shown in Table 4. Our manual analysis shows that all 5 authors are the most influential researchers in the corresponding research field. Because it is difficult to evaluate the true popularity of users in Reddit, we do not show the case on Reddit.

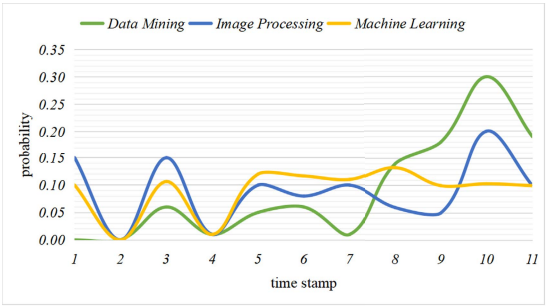

FIGURE 9. Temporal-topic variation of a user.

5) TEMPORAL VARIATION OF USER TOPICS

Users in a community share similar topics, which generates community topics. However, users' interested topics are changing over time, which is modeled by the fourth social behavior. This case study investigates the changing of topics on individual level.

Due to the large number of users, Fig. 9 shows only the topic changes of the first author in the community *Image Processing*. We can track the change in his research topics, which is significant for better understanding the change in the leading research direction. The DBLP dataset does not include publications at time stamp 2 and time stamp 4. Fig. 9 shows that his studies on *Image Processing* and *Machine Learning* are stable except for the above time stamps, and his publications on *Data Mining* increased at the last four time stamps, i.e., 8, 9, 10, and 11.

VI. CONCLUSION AND DISCUSSION

First, we investigated and assessed the influence and importance of considering user social behaviors for community detection. Social behaviors exhibit users' habits and make significant contributions to the interactions among users (i.e., the topology structure and content of a network). We proposed four types of user social behaviors (i.e., reciprocity of interactions, posting preference, multitopic preference, and temporal variation of topics). Second, we proposed a novel method (SBCD) by combining user social behaviors, network topology and network content seamlessly in a generative model. It investigates the formation of a network with complex content to infer community structure and community topics. In this model, network content is divided into node content and link content. Third, we evaluate SBCD on two real networks with ground truths and compare it with six state-of-the-art methods. The experimental results show that SBCD improves the accuracy of community detection by considering social behaviors. Finally, SBCD can also identify topics, topic distributions of communities, user popularity, and individual topic changes over time in each community. In the future, we first intend to get a proper synthetic benchmark to evaluate our community detection model. There are two key requirements for synthetic benchmark. (1) The social networks in our paper should include textual content on both nodes and links, which means that there should be a word

set corresponding to different topics for each node and each link. The words in the set should be from a meaningful true post. The first three behaviors can be satisfied by existing benchmark generation methods; (2) Benchmark generation methods should generate temporal topics of nodes which should evolve in the similar way as reality. Second, we intend to investigate how community members and community topics evolve as a result of the changing of users' topics.

ACKNOWLEDGMENT

Yingkui Wang thanks to the assistance of Prof. Jianxia Bai from Tianjin Renai College.

REFERENCES

[1] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.
[2] H. Roghani and A. Bouyer, "A fast local balanced label diffusion algorithm for community detection in social networks," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 25, 2022, doi: 10.1109/TKDE.2022.3162161.
[3] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, "Overlapping community detection using Bayesian non-negative matrix factorization," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 83, no. 6, Jun. 2011, Art. no. 066114.
[4] 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, pp. 1–35, Aug. 2013.
[5] 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, Apr. 2002.
[6] D. Jin, B. Li, P. Jiao, D. He, H. Shan, and W. Zhang, "Modeling with node popularities for autonomous overlapping community detection," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–23, Jun. 2020.
[7] D. He, Y. Song, D. Jin, Z. Feng, B. Zhang, Z. Yu, and W. Zhang, "Community-centric graph convolutional network for unsupervised community detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3515–3521.
[8] D. Jin, B. Zhang, Y. Song, D. He, Z. Feng, S. Chen, W. Li, and K. Musial, "ModMRF: A modularity-based Markov random field method for community detection," *Neurocomputing*, vol. 405, pp. 218–228, Sep. 2020.
[9] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, P. Yu, and W. Zhang, "A survey of community detection approaches: From statistical modeling to deep learning," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 11, 2021, doi: 10.1109/TKDE.2021.3104155.
[10] 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.
[11] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, "An introduction to exponential random graph (p*) models for social networks," *Soc. Netw.*, vol. 29, no. 2, pp. 173–191, 2007.
[12] 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.
[13] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3, pp. 75–174, Jan. 2010.
[14] D. He, W. Song, D. Jin, Z. Feng, and Y. Huang, "An end-to-end community detection model: Integrating LDA into Markov random field via factor graph," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5730–5736.
[15] 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. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 116–124.
[16] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.
[17] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 1151–1156.
[18] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociol.*, vol. 27, no. 1, pp. 415–444, Aug. 2001.
[19] P. Chunaev, T. Gradov, and K. Bochenina, "Community detection in node-attributed social networks: How structure-attributes correlation affects clustering quality," *Proc. Comput. Sci.*, vol. 178, pp. 355–364, Jan. 2020.
[20] P. Chunaev, "Community detection in node-attributed social networks: A survey," *Comput. Sci. Rev.*, vol. 37, Aug. 2020, Art. no. 100286.
[21] P. Chunaev, I. Nuzhdenko, and K. Bochenina, "Community detection in attributed social networks: A unified weight-based model and its regimes," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 455–464.
[22] K. Musiał and P. Kazienko, "Social networks on the internet," *World Wide Web*, vol. 16, no. 1, pp. 31–72, 2013.
[23] D. Jin, X. Wang, M. Liu, J. Wei, W. Lu, and F. Fogelman-Soulie, "Identification of generalized semantic communities in large social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2966–2979, Oct. 2020.
[24] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 1, pp. 1–28, Feb. 2014.
[25] 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.
[26] S. Pool, F. Bonchi, and M. V. Leeuwen, "Description-driven community detection," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 2, pp. 1–28, 2014.
[27] D. Jin, X. Wang, D. He, J. Dang, and W. Zhang, "Robust detection of link communities with summary description in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2737–2749, Jun. 2019.
[28] 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.
[29] C. R. Shalizi and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies," *Sociol. Methods Res.*, vol. 40, no. 2, pp. 211–239, 2010.
[30] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 631–640.
[31] M. Wang, C. Wang, J. X. Yu, and J. Zhang, "Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework," *Proc. VLDB Endowment*, vol. 8, no. 10, pp. 998–1009, 2015.
[32] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.
[33] M. Chen, T. Nguyen, and B. K. Szymanski, "A new metric for quality of network community structure," *Comput. Sci.*, vol. 4, no. 2, pp. 22–29, 2015.
[34] Y. Zhu, X. Yan, L. Getoor, and C. Moore, "Scalable text and link analysis with mixed-topic link models," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 473–481.
[35] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 542–550.
[36] J. Chang and D. Blei, "Relational topic models for document networks," in *Proc. Artif. Intell. Statist. (AISTATS)*, vol. 5. Clearwater Beach, FL, USA: W&CP, 2009, pp. 81–88.
[37] S. Li, T.-S. Chua, J. Zhu, and C. Miao, "Generative topic embedding: A continuous representation of documents," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 666–675.
[38] X. Li, C. Li, J. Chi, J. Ouyang, and C. Li, "Dataless text classification: A topic modeling approach with document manifold," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 973–982.
[39] D. Jin, K. Wang, G. Zhang, P. Jiao, D. He, F. Fogelman-Soulie, and X. Huang, "Detecting communities with multiplex semantics by distinguishing background, general, and specialized topics," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 11, pp. 2144–2158, Nov. 2020.
[40] I. Falih, N. Grozavu, R. Kanawati, and Y. Bennani, "Community detection in attributed network," in *Proc. Companion Web Conf. (WWW)*, 2018, pp. 1299–1306.
[41] S. Citraro and G. Rossetti, "X-mark: A benchmark for node-attributed community discovery algorithms," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–14, Dec. 2021.
[42] M. Sachan, A. Dubey, S. Srivastava, E. P. Xing, and E. Hovy, "Spatial compactness meets topical consistency: Jointly modeling links and content for community detection," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, Feb. 2014, pp. 503–512.

[43] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "A model-based approach to attributed graph clustering," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2012, pp. 505–516.
[44] B. Guidi, M. Conti, A. Passarella, and L. Ricci, "Managing social contents in decentralized online social networks: A survey," *Online Social Netw. Media*, vol. 7, pp. 12–29, Sep. 2018.
[45] B. Guidi, A. Michienzi, and G. Rossetti, "Towards the dynamic community discovery in decentralized online social networks," *J. Grid Comput.*, vol. 17, no. 1, pp. 23–44, Mar. 2019.
[46] B. Guidi, A. Michienzi, and L. Ricci, "SONIC-MAN: A distributed protocol for dynamic community detection and management," in *Proc. IFIP Int. Conf. Distrib. Appl. Interoperable Syst.* Madrid, Spain: Springer, 2018, pp. 93–109.
[47] B. Guidi, A. Michienzi, and G. Rossetti, "Dynamic community analysis in decentralized online social networks," in *Proc. Eur. Conf. Parallel Process.* Santiago de Compostela, Spain: Springer, 2017, pp. 517–528.
[48] Z. Hu, J. Yao, B. Cui, and E. Xing, "Community level diffusion extraction," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2015, pp. 1555–1569.
[49] 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.
[50] Y. Yang, J. Tang, C. W.-K. Leung, Y. Sun, Q. Chen, J. Li, and A. Q. Yang, "Rain: Social role-aware information diffusion," in *Proc. 29th AAAI Conf. Artif. Intell.*, vol. 15, 2015, pp. 367–373.
[51] X. Wang, D. Jin, M. Liu, D. He, K. Musial, and J. Dang, "Emotional contagion-based social sentiment mining in social networks by introducing network communities," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1763–1772.
[52] Y. Han and J. Tang, "Probabilistic community and role model for social networks," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 407–416.
[53] X. Wang, X. Cao, D. Jin, Y. Cao, and D. He, "The (un)supervised NMF methods for discovering overlapping communities as well as hubs and outliers in networks," *Phys. A, Stat. Mech. Appl.*, vol. 446, pp. 22–34, Mar. 2016.
[54] E. Zhong, W. Fan, J. Wang, L. Xiao, and Y. Li, "ComSoc: Adaptive transfer of user behaviors over composite social network," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 696–704.
[55] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2010, pp. 1633–1636.
[56] A. De Salve, M. Dondio, B. Guidi, and L. Ricci, "The impact of user's availability on on-line ego networks: A Facebook analysis," *Comput. Commun.*, vol. 73, pp. 211–218, Jan. 2016.
[57] B. Guidi, A. Michienzi, and L. Ricci, "Managing communities in decentralised social environments," *Peer-to-Peer Netw. Appl.*, vol. 2022, pp. 1–26, Jul. 2022.
[58] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya–Gamma latent variables," *J. Amer. Stat. Assoc.*, vol. 108, no. 504, pp. 1339–1349, Dec. 2013.
[59] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.
[60] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *Proc. KDD*, 2008, pp. 990–998.
[61] J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang, "Scalable inference for logistic-normal topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2445–2453.
[62] 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.
[63] 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, pp. 321–328, 2019.
[64] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for K-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2009, pp. 877–886.
[65] Q. Diao, J. Jiang, F. Zhu, and E. P. LIM, "Finding bursty topics from microblogs," in *Proc. ACL*, 2012, pp. 1–10.
[66] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, "A unified model for stable and temporal topic detection from social media data," in *Proc. IEEE 29th Int. Conf. Data Eng. (ICDE)*, Apr. 2013, pp. 661–672.

YINGKUI WANG received the B.E. and M.E. degrees in statistics from Sichuan University, Sichuan, China, in 2005 and 2008, respectively, and the Ph.D. degree from Tianjin University, China, in 2021. He is a Lecturer with the School of Computer Science and Technology, Tianjin Renai Colledge. His current research interests include community detection and deep learning.

DI JIN received the B.S., M.S., and Ph.D. degrees in computer science from Jilin University, Changchun, China, in 2005, 2008, and 2012, respectively. He was a Postdoctoral Research Fellow at the School of Design, Engineering, and Computing, Bournemouth University, Poole, U.K., from 2013 to 2014. He is an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. He has published over 40 international journal articles and conference papers. His current research interests include data mining, complex network analysis, and machine learning.

DONGXIAO HE received the Ph.D. degree in computer science from Jilin University, Changchun, China, in 2014. She was a Postdoctoral Research Fellow at the Department of Computer Science, Dresden University of Technology, Germany, from 2014 to 2015. She is an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. She has published over 50 international journal and conference papers. Her current research interests include data mining and analysis of complex networks.

KATARZYNA MUSIAL was born in Poland, in 1982. She received the M.Sc. degree in computer science from the Wroclaw University of Technology, Poland, the second M.Sc. degree in software engineering from the Blekinge Institute of Technology, Sweden, in 2006, and the Ph.D. degree from the Institute of Informatics, Wroclaw University of Technology, in 2009. She focused her Ph.D. thesis on the calculation of individual's social position in the virtual social network. In September 2017, she joined as Associate Professor of network science with University of Technology Sydney (UTS). She is interested especially in complex social networks and dynamics and evolution of complex networked systems.

JIANWU DANG (Member, IEEE) received the graduate and M.S. degrees from Tsinghua University, China, in 1982 and 1984, respectively, and the Ph.D. degree from Shizuoka University, Japan in 1992. He was a Lecturer at Tianjin University, from 1984 to 1988. He was a Senior Researcher at the ATR Human Information Processing Laboratories, Japan, from 1992 to 2001. He joined the University of Waterloo, Canada, as a Visiting Scholar for one year from 1998. Since 2001, he has been with the Japan Advanced Institute of Science and Technology (JAIST). He joined the Institute of Communication Parlee (ICP), Center of National Research Scientific, France, as a research scientist the first class, from 2002 to 2003. His research interests include speech production, speech synthesis, and speech cognition. He built a 3D physiological model for speech and swallowing, and endeavors to apply the model on clinics.