IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 1, 2020, accepted June 21, 2020, date of publication June 30, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3005935

Community Detection Based on Individual Topics and Network Topology in Social Networks

HUI JIANG[1], LINJUAN SUN[1], JUAN RAN[1], JIANXIA BAI[2], AND XIAOYE YANG[2]
[1]Department of Computer Science and Technology, Tianjin University Renai College, Tianjin 300072, China
[2]Department of Mathematics, Tianjin University Renai College, Tianjin 300072, China

Corresponding author: Hui Jiang (jianghui_nudt@163.com)

This work was supported by the Scientific Research Project of Tianjin Municipal Education Commission under Grant 2018KJ267.

ABSTRACT Detecting community structures is an important research topic in social network analysis. Unfortunately, the fundamental factors that drive the generation of social networks (i.e., the network topology and content) and community structures have not been well investigated. In this paper, according to the natural characteristics of social networks, we reveal that individual topics play a core role in community generation. If two individuals are in the same community and are interested in similar topics, it is more likely that a link will form between them. Otherwise, the probability of generating a link depends on the relationships between their communities and the topics they talk about. Based on the above observations, a novel generative community detection model is proposed that simulates the generation of the network topology and network content by considering individual topics. Moreover, our model utilizes a topic model to generate network content. The model is evaluated on two real-world datasets. The experimental results show that the community detection results outperform all the state-of-the-art baselines. In addition to accurate community detection results, we identify each individual topic distribution and the most popular users corresponding to different topics in each community.

INDEX TERMS Community detection, individual topics, social networks.

I. INTRODUCTION

Studies on complex networks [1], [2] have become increasingly important. In this paper, we focus on a special type of social network. This type not only include complex topology but also contains rich text content, e.g., Reddit, Twitter, and co-authorship networks. Taking Reddit as an example, when user *j* replies to a post of user *i*, a directed link is generated from *j* to *i*, and the content that user *j* replies to is used as link content. In this type of network, all links are associated with text content.

In the research on social networks, community detection has been a hot topic in recent years [3]–[5]. A community is a group of individuals that are connected densely, while individuals from different groups are connected sparsely [6]. The community structure is an innate characteristic of a social network; therefore, it is important to detect communities for a better understanding of the compositions of social networks.

Recently, a large number of community detection methods have been proposed [7], [8]. Earlier studies only focused on network topology, and they ignored network content, which includes node content and link content because specific networks do not supply content information, e.g., neural networks and computer networks. However, for networks with content, e.g., Twitter, Facebook, and Reddit, content information is useful for community detection. Network content is generated by nodes to express ideas about various topics. Specifically, link content reflects what users want to communicate about. The willingness to interact drives the generation of links between users. Therefore, the generation of a network topology that a community structure forms is related to the link content. Moreover, network content supplies the semantic information of communities, i.e., topic distributions. Identifying community semantics (e.g., community-topic distributions) is important for better understanding communities. Many models have been proposed to detect community structures and community semantics at the same time [9]. How to integrate network content and network topology seamlessly is a challenge [10].

Although many methods that utilize network topology and network content have been proposed, they have a common drawback. The factors that drive the generation of community

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

structures have not been studied. In the generation process of network topology and network content, a community structure forms under the influence of some key factors. Investigating these factors is important because of the following three issues.

First, revealing the driving factors is conducive to better understanding the innate characteristics of social networks, which explains how the topology is generated. Second, the underlying community generation mechanism is unknown. Investigating the driving factors can help derive a precise community structure and observe the mechanism that forms the community structure, which is important for further community evolution tracking. Finally, beyond the community structure, mining community semantics is also important for understanding communities in networks with content. Network content reflects the inner semantic profile of networks. Therefore, observing the factors that drive the formation of networks and community structure can help us better understand the process of community generation and identify accurate community structures with semantic information.

To achieve objectives in the above issues, we analyse a large number of networks and find that individual topics play a significant role in driving the generation of network topologies, network content, community structures, and community semantics. As discussed above, each link includes text content in our networks of interest, i.e., the link content. Link content reflects what the source node wants to talk about with the target node of a link. The link itself implies that the tendency of interaction from the source node to the target node is high. The significance of individual topics in social networks is explained as follows.

First, we consider the impacts of individual topics on the generation of network topologies. If two users in a network share similar topic preferences, then the probability of forming a link between them is large. Otherwise, the probability is small. If all the users keep their topics of interest unchanged, then the network topology would be stable. If a user changes topics of interest at some time point, then he or she will communicate with users who are also interested in these topics. In this way, new links form between a new pair of users. In summary, the change in individual topics affects the network topology.

Second, network content is generated by individuals. Since a community is a set of users, to identify community semantics, the topics discussed by individuals inside a community need to be analysed. Community semantics are the result of the aggregation of individual topics. Therefore, focusing on individual-level topic preferences can help us understand the generation of community-level topic distributions.

Third, in social network analysis, there is a gap between individual-level research and network-level research. The study of individuals is limited to a specific node, so it cannot reflect the characteristics of the networks. Network-level studies ignore the individual contributions to the characteristics of networks. The community structure is an efficient solution to fill this gap. The community structure mainly focuses on the characteristics of communities, i.e., community members, community semantics and the diffusion between communities (semantics/diffusion). A community is a group of individuals. Unfortunately, existing community models cannot explain the procedure of generating communities from individuals. They only detect community results but do not reveal the reasons in a more fine-grained manner.

Finally, considering the links and semantics in a network, the definition of a community leads to a conflict between these concepts, which means that individuals in the same community interact with each other densely, but they also share different topics of discussion [9]. Since a community is a high-order structure, there might be no links between individuals even if they are from the same community. Therefore, in this paper, we assume that even if two users are in the same community, the probability of generating interactions between them is decided by their topics of interest. Unlike previous works that assume a fixed probability of forming a link between any two users from the same community or from different communities, we consider the individual topic preference both inside communities and among communities. For those users who are in different communities, if they share similar topics of interest, they might still interact with each other and generate inter-community links. Moreover, our idea reflects an important phenomenon that there are correlations between topics. Users interested in highly correlated topics are more likely to generate links no mater they are in the same community or not.

Based on the above discussions, we propose a novel generative model: CDITN (Community Detection based on Individual Topics and Network topology). It integrates individual topics, network topology and network content seamlessly. It is composed of two components in which the network topology and network content originate from individual topics. In the formation process, the model discriminates the community distributions of the source node and target node on a link. It indicates that the source node and target node might be in different communities even if an interaction between them occurs. The probability of forming a link depends on the consistency in their topics of interest, which means that the probability of interactions is large if two nodes share the same topics of interest. Furthermore, studies have shown that there are different diffusions between communities [11]. We summarize our contributions as follows:

- We, for the first time, investigate a key driving factor of community generation, i.e., individual-level topics.
- We propose a novel generative model to generate a network topology and network contents. Moreover, the individual topic distribution and most important users corresponding to different topics in each community are detected.
- We conduct experiments on two datasets. The results show that CDITN outperforms all the baseline models.

The rest of the paper is organized as follows: Section 2 reviews related works on community detection; Section 3

TABLE 1. Notations.

Notation	Description
U	user set
K	topic set
C	community set
V	vocabulary
E_i	link set of user i
e_{ij}	a directed link from user i to ij
W_{ir}	word list of e_{ij}
W_{irq}	the q-th word of W_{ir}
$\boldsymbol{\pi}_i$	multinomial distribution over communities of user i
$\boldsymbol{\theta}_i$	multinomial distribution over topics of user i
$\boldsymbol{\phi}_k$	multinomial distribution over words of topic k
$\boldsymbol{\eta}_c$	multinomial distribution over communities of community c
$\boldsymbol{\xi}_{ck}$	multinomial distribution over users of community c and topic k
c_{ir}	community indicator of user i for e_{ij}
c'_{ij}	community indicator of user j for e_{ij}
y_{ir}	topic indicator of the r-th link of user i
$\alpha, \lambda, \beta, \varepsilon, \rho$	Dirichlet priors

describes our model in detail; Section 4 describes the model inference method; 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, we review related work for community detection. Specifically, we discuss similar studies that consider network content.

In recent years, a large number of models for community detection have been proposed [12]–[14]. From the view of what type of data is utilized, these models can be classified into two categories. The first category of models is based on network topology, e.g., Louvain [15], LPA [16] algorithms. The second category of models integrates the network topology and network content into a unified model [17]–[19].

Earlier community detection models only use the topology information, e.g., spectral clustering [20], hierarchical clustering [6], and modularity-based methods [15]. They achieve accurate community detection results in some networks with clear topological structures. Unfortunately, when the network structure is complex in the real world, their performance is limited. Thus, topology information alone is not sufficient for accurately detecting the community structure.

Recently, an increasing number of studies have integrated network topology and network content for community detection [21]–[23]. Utilizing network content not only improves the accuracy of community detection results but also gives us the semantic description of communities [24]–[26]. Moreover, [11] investigated topic-diffusion patterns across communities. Reference [27], for the first time, proposed the concept of community profiling and considered two types of links: friendship links and diffusion links. References [11] and [27] proposed generative models in which the network topology and network content were generated. Zhongying Zhao *et al.* proposed a novel incremental method to detect communities in dynamic evolving social networks. The method analyzes incremental dynamic changes and updates community structure incrementally. It explains how communities evolve. While, our method explains the generation of communities from individual level. In addition to detecting the community structure, the method proposed by [28] provided two-level semantic interpretation for each community, i.e., general topics and specialized topics. The extension of [28] proposed a model to detect communities with multiplex semantics by distinguishing background, general and specialized Topics. Reference [29] explored the intrinsic correlation between communities and topics to discover link communities. Their method extracted community summaries in sentences instead of words for topic labelling. Reference [30] investigated topic correlations in social networks, which affect link generation between nodes.

However, the key factors that drive the generation of communities are not well explained. The factors first affect the generation of networks, including the topology and content. During the generation process, the community structure forms. Revealing the key factors can help us better understand the precise semantic information of communities.

III. NOTATION AND MODEL

In this section, we first introduce notations used in our model. Then, we formulate the task of community detection by utilizing individual topics. Thereafter, a novel generative model is proposed, i.e., CDITN. Finally, its process of generating network topology and network content is presented.

A. NOTATIONS

The notations used in our model are shown in Table 1. We describe all the definitions as follows.

Definition 1. A network consists of three components: $G = (U, E, W)$. U denotes the user set. E is the link set. W denotes the set of link content.

Definition 2. For user i, his or her community distribution is defined by a vector $\boldsymbol{\pi}_i$. The dimension of user i is the number of communities, which follows a multinomial distribution over all the communities. The probability of user i belonging to community c is $\boldsymbol{\pi}_{ic}$. In this paper, we set a threshold to determine the community of each user. If $\boldsymbol{\pi}_{ic}$ is larger than the threshold, user i belongs to community c.

Definition 3. A topic k is defined by a vector $\boldsymbol{\phi}$. $\boldsymbol{\phi}$ follows a multinomial distribution, and the dimension is the number of vocabularies. We fix the vocabulary of the corpus to a fixed set denoted by V. The probability of a word w belonging to topic k is denoted by $\boldsymbol{\phi}_{kw}$.

Definition 4. Individual topic interests are defined by a vector $\boldsymbol{\theta}_i$ for user i. $\boldsymbol{\theta}_i$ follows a multinomial distribution, and the dimension is the number of topics, i.e., $|K|$. A user might be interested in various topics. The probability of user i being interested in topic k is denoted by $\boldsymbol{\theta}_{ik}$.

Definition 5. Interactions between communities are defined by a vector $\boldsymbol{\eta}$. $\boldsymbol{\eta}$ follows a multinomial distribution, and the dimension is the number of communities. For community c, the probability of its interaction with community c' is $\boldsymbol{\eta}_{c,c'}$. Interactions between communities always exist, but they interact with each other with different probabilities. Take the co-authorship network as an example. There are three communities: *machine learning*, *image processing*, and *software engineering*. The interactions between the *machine learning* and *image processing* communities are denser than the other pairs of communities.

Definition 6. User popularity in each community corresponding to topics is defined by a vector $\boldsymbol{\xi}$ that follows a multinomial distribution, and the dimension is the number of users. For community c and topic k, the probability of its users being selected by others to communicate with is denoted by $\boldsymbol{\xi}_{ck,i}$. The key idea is that the probability of interaction between users is related to the two users' communities and the topics they are talking about.

B. MODEL FORMULIZATION

In this section, we first formulate the problem to be resolved. Then, we describe our model in detail.

Based on the above definitions, the problem we need to solve is described as follows. Given a network with content as input, for each user, we need to infer his or her community membership, topic distribution and popularity in each community corresponding to each topic. For each community, we need to infer the probability of the user's interaction with all communities.

We propose a novel generative model to generate all the observed data, i.e., the network topology and content from the parameters and latent variables. The graphical model is shown in Fig. 1. The observed data include all the links and link content, i.e., the directed link e_{ij} for all users and

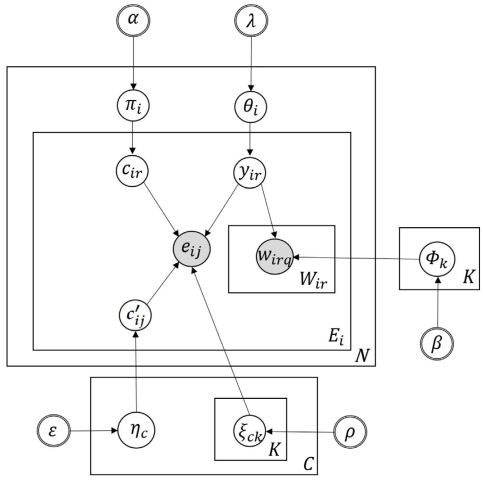


FIGURE 1. Graphical model.

link content $W_{e_{ij}}$. These data are generated by the following process.

For each directed link e_{ij} from source user i to target user j, we first sample user i's community indicator c_{ir} from his or her community membership distribution $\boldsymbol{\pi}_i$. c_{ir} is a latent variable. Second, we sample the topic indicator y_{ir} from user i's topic distribution. This indicator is a latent variable and denotes the topic of link e_{ij}. Third, we need to estimate which community user j is most likely from. Thereafter, community indicator c'_{ij} is sampled from $\boldsymbol{\eta}_{c_{ir}}$, where c_{ir} is the community of user i sampled at the previous step. Finally, based on community indicator c'_{ij} and topic indicator y_{ir}, we sample user j from $\boldsymbol{\xi}_{c'_{ij},y_{ir}}$. In this way, link e_{ij} is generated.

For link content W_{ir}, we need to sample all the words, which is denoted by w_{irq}. We use a method such as Twitter-LDA [31]. Based on the topic indicator y_{ir} of the current link and word distribution of $\boldsymbol{\phi}_{y_{ir}}$, each word is sampled.

The key innovation of our model is the adoption of individual topic distribution $\boldsymbol{\theta}_i$, which promotes investigations on semantics on a smaller scale, i.e., the individual level. The generation of network topology and network content is derived from elementary factors. Moreover, the relations between communities and links are modelled. The generative process of our model is summarized as follows.

1) For each topic y,
   a) Sample its word distribution from a Dirichlet prior: $\boldsymbol{\phi}_k \mid \beta \sim Dir(\beta)$;
2) For each community c,
   a) Sample the community distribution from a Dirichlet prior: $\boldsymbol{\eta}_c \mid \varepsilon \sim Dir(\varepsilon)$;
   b) For each topic k,
      i) Sample the user distribution from a Dirichlet prior: $\boldsymbol{\xi}_{ck} \mid \rho \sim Dir(\rho)$;
3) For each user i in U,
   a) Sample the community distribution from a Dirichlet prior: $\boldsymbol{\pi}_i \mid \alpha \sim Dir(\alpha)$;
   b) Sample the individual topic distribution from a Dirichlet prior: $\boldsymbol{\theta}_i \mid \lambda \sim Dir(\lambda)$;

4) For each user i in U,
   a) For each directed link e_{ir} in E_i,
      i) Sample the community of node i from a Multinomial distribution: $c_{ir} \mid \boldsymbol{\pi}_i \sim Mul(\boldsymbol{\pi}_i)$;
      ii) Sample topic y_{ir} from a Multinomial distribution: $y_{ir} \mid \boldsymbol{\theta}_i \sim Mul(\boldsymbol{\theta}_i)$.
      iii) Sample the community of node j (target node of e_{ir}) c'_{ij} from a Multinomial distribution: $c'_{ij} \mid \boldsymbol{\eta}_{c_{ir}} \sim Mul(\boldsymbol{\eta}_{c^i_{ir}})$;
      iv) Sample node j from a Multinomial distribution: $e_{ij} \mid \boldsymbol{\xi}_{c'_{ij}y_{ir}} \sim Mul(\boldsymbol{\xi}_{c'_{ij}y_{ir}})$
      v) For each word w_{irq} in W_{ir},
         • Sample the word from a Multinomial distribution: $w_{irq} \mid \boldsymbol{\phi}_{y_{ir}} \sim Mul(\boldsymbol{\phi}_{y_{ir}})$;

IV. MODEL INFERENCE

In this section, we estimate all parameters, i.e., $\boldsymbol{\pi}$, $\boldsymbol{\theta}$,$\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\phi}$.

A. APPROXIMATE INFERENCE

Based on the probabilistic graphical model, we first obtain the posterior distribution shown by (1).

$$\begin{aligned} &P(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\phi}, c, y, c' | U, E, W, \alpha, \lambda, \beta, \varepsilon, \rho) \\ &\propto P(c|\boldsymbol{\pi})P(\boldsymbol{\pi}|\alpha) \\ &\quad \cdot P(y|\boldsymbol{\theta})P(\boldsymbol{\theta}|\lambda) \\ &\quad \cdot P(c'|\boldsymbol{\eta}, c)P(\boldsymbol{\eta}|\varepsilon) \\ &\quad \cdot P(e|\boldsymbol{\xi}, c, c', y)P(\boldsymbol{\xi}|\rho) \\ &\quad \cdot P(\boldsymbol{\phi}|\beta)P(w|\boldsymbol{\phi}, y). \end{aligned} \tag{1}$$

where the user set U, link set E, and link content set W are observed data. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\theta}$,$\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\phi}$ are to be estimated. Because calculating the normalizing constant is hard, we adopt collapsed Gibbs Sampling to solve the inference problem.

Based on (1), we marginalize all parameters, i.e., $\boldsymbol{\pi}$, $\boldsymbol{\theta}$,$\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\phi}$. We obtain (2):

$$\begin{aligned} &P(c, c', y|.) \\ &\propto \int P(\boldsymbol{\pi}|\alpha)P(c|\boldsymbol{\pi})d\boldsymbol{\pi} \\ &\quad \cdot \int P(y|\boldsymbol{\theta})P(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} \\ &\quad \cdot \int P(c'|\boldsymbol{\eta}, c)P(\boldsymbol{\eta}|\varepsilon)d\boldsymbol{\eta} \\ &\quad \cdot \int P(e|\boldsymbol{\xi}, c', y)P(\boldsymbol{\xi}|\rho)d\boldsymbol{\xi} \\ &\quad \cdot \int P(\boldsymbol{\phi}|\beta)P(w|\boldsymbol{\phi}, y)d\boldsymbol{\phi}. \end{aligned} \tag{2}$$

The next step is to calculate all the integrals in (2). The first integral is calculated by the following equation.

$$\int P(\boldsymbol{\pi}|\alpha)P(c|\boldsymbol{\pi})d\boldsymbol{\pi} = \prod_i^{|U|} \frac{\Gamma(|C|\alpha)}{(\Gamma(\alpha_i))^{|C|}} \cdot \frac{\prod_c^{|C|} \Gamma(n_i^{(c)} + \alpha)}{\Gamma(n_i^{(\cdot)} + |C|\alpha)}, \tag{3}$$

where $n_i^{(c)}$ is the number of links assigned to community c for user i. $n_i^{(\cdot)}$ is the total number of links that are assigned to all communities for user i. For the rest of this paper, dots in parentheses denote marginalizing out all the latent variables, e.g., the community indicator c and the topic indicator k.

The second integral in (2) is calculated by (4).

$$\int P(y|\boldsymbol{\theta})P(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} = \frac{\Gamma(|K|\delta)}{\Gamma(\delta)^{|K|}} \cdot \prod_i^{|U|} \frac{\prod_k^{|K|} \Gamma(n_i^{(k)} + \delta)}{\Gamma(n_i^{(\cdot)} + |K|\delta)}, \tag{4}$$

where $n_i^{(}k)$ is the number of links assigned to topic k for user i.

The third integral in (2) is calculated by (5).

$$\int P(c'|\boldsymbol{\eta}, c)P(\boldsymbol{\eta}|\varepsilon)d\boldsymbol{\eta} = \prod_c^{|C|} \frac{\Gamma(|C|\varepsilon)}{(\Gamma(\varepsilon))^{|C|}} \cdot \frac{\prod_m^{|C|} \Gamma(n_c^{(m)} + \varepsilon)}{\Gamma(n_c^{(\cdot)} + |C|\varepsilon)}, \tag{5}$$

where $n_c^{(m)}$ denotes the number of links whose source node is in community c and whose target node is in community m.

The fourth integral in (2) is calculated by (6).

$$\int P(e|\boldsymbol{\xi}, c', y)P(\boldsymbol{\xi}|\rho)d\boldsymbol{\xi} = \prod_i^{|U|} \prod_e^{|E_i|} \frac{\Gamma(|U|\rho)}{(\Gamma(\rho))^{|U|}} \cdot \frac{\prod_u^{|U|} \Gamma(n_{c'y}^{(u)} + \rho)}{\Gamma(n_{c'y}^{(\cdot)} + |U|\rho)}, \tag{6}$$

where $n_{c'y}^{(u)}$ is the number of links with target node u in community c' and for topic y.

The third integral in (2) is calculated by (7).

$$\int P(\boldsymbol{\phi}|\beta)P(w|\boldsymbol{\phi}, y)d\boldsymbol{\phi} = \prod_k^{|K|} \frac{\Gamma(|V|\beta)}{(\Gamma(\beta))^{|V|}} \cdot \frac{\prod_w^{|V|} \Gamma(n_y^{(w)} + \beta)}{\Gamma(n_y^{(\cdot)} + |V|\beta)}, \tag{7}$$

where $n_y^{(w)}$ denotes the number of times that word w is assigned to topic y.

Aggregating all the integrals calculated above, we further sample all the latent variables. For each link of user i, the community indicator is sampled by (8)

$$\begin{aligned} P(c_{ir} &= c|c_{\neg ir}, y_{ir} = k, c' = m, .) \\ &= \frac{P(c, y, c'|.)}{P(c_{\neg ir}, y, c'|.)} \\ &= \frac{n_{i,\neg ir}^{(c)} + \alpha}{n_{i,\neg ir}^{(\cdot)} + |C|\alpha} \cdot \frac{n_{ck,\neg ir}^{(j)} + \rho}{n_{ck,\neg ir}^{(\cdot)} + |U|\rho} \cdot \frac{n_{c,\neg ir}^{(m)} + \varepsilon}{n_{c,\neg ir}^{(\cdot)} + |C|\varepsilon}, \end{aligned} \tag{8}$$

where $n_{i,\neg ir}^{(c)}$ is the number of links assigned to community c for user i, excluding link e_{ij}. $n_{ck,\neg ir}^{(j)}$ denotes the number of links whose target node j is in community c and with topic k, excluding link e_{ij}. $n_{c,\neg ir}^{(m)}$ is the number of links whose source node is in community c and whose target node is in community m, excluding link e_{ij}.

For each link e_{ij}, the community indicator of the target node is sampled by (9)

$$\begin{aligned} P(c'_{ij} &= c' | c'_{\neg ij}, y_{ir} = k, c = m, .) \\ &= \frac{P(c, y, c' | .)}{P(c'_{\neg ij}, y, c | .)} \\ &= \frac{n^{(j)}_{c'k,\neg ij} + \rho}{n^{(\cdot)}_{c'k,\neg ij} + |U|\rho} \cdot \frac{n^{(c')}_{m,\neg ij} + \varepsilon}{n^{(\cdot)}_{m,\neg ij} + |C|\varepsilon}, \end{aligned} \tag{9}$$

where $n^{(j)}_{c'k,\neg ij}$ and $n^{(c')}_{m,\neg ij}$ represent the same values as in (9). The topic of each link is sampled following (10).

$$\begin{aligned} P(y_{ir} &= k | y_{\neg ir}, c_{ir} = c, c') = m, .) \\ &= \frac{P(c, y, c' | .)}{P(y_{\neg ir}, c, c' | .)} \\ &= \frac{n^{(k)}_{i,\neg ir} + \lambda}{n^{(\cdot)}_{i,\neg ir} + |K|\lambda} \cdot \frac{\prod_{v}^{|V|} \prod_{s=0}^{n^{(v)}_{ir}-1} (n^{(v)}_{k,\neg ir} + s + \beta)}{\prod_{s=0}^{n^{(\cdot)}_{ir}-1} (n^{(\cdot)}_{k,\neg ir} + s + \beta)} \\ &\quad \cdot \frac{n^{(j)}_{c'k,\neg ir} + \rho}{n^{(\cdot)}_{c'k,\neg ir} + |U|\rho}, \end{aligned} \tag{10}$$

where $n^{(v)}_{ir}$ is the number of times that word v appears in the link content of e_{ij}. $n^{(v)}_{k,\neg ij}$ is the number of times that word v is assigned to topic k, excluding e_{ij}.

B. PARAMETER ESTIMATION

Finally, we estimate all the parameters based on the samples derived by running the Gibbs sampler until convergence.

$$\hat{\boldsymbol{\pi}}_{ic} = \frac{n^{(c)}_{i} + \alpha}{n^{(\cdot)}_{i} + |C|\alpha}. \tag{11}$$

$$\hat{\boldsymbol{\xi}}_{ck,i} = \frac{n^{(i)}_{ck} + \rho}{n^{(\cdot)}_{ck} + |U|\rho}. \tag{12}$$

$$\hat{\boldsymbol{\phi}}_{kv} = \frac{n^{(v)}_{k} + \beta}{n^{(\cdot)}_{k} + |V|\beta}. \tag{13}$$

$$\hat{\boldsymbol{\eta}}_{cg} = \frac{n^{(g)}_{c} + \varepsilon}{n^{(\cdot)}_{c} + |C|\varepsilon}. \tag{14}$$

$$\hat{\boldsymbol{\theta}}_{ik} = \frac{n^{(k)}_{i} + \lambda}{n^{(\cdot)}_{i} + |K|\lambda}. \tag{15}$$

C. TIME COMPLEXITY ANALYSIS

The pseudo-code of our model is illustrated in Alg. 1. Then, we analyse the time complexity of our algorithm. In the above algorithm, the number of topics and communities are fixed to the true value according to the ground truth, i.e., $|K|$ and $|C|$. Steps 5-7 sample a community with two nodes and the topic for each link. In (8) and (9), all the counters are stored in memory; therefore, they take a constant time for a link. To calculate (10), there are $|V|$ words; therefore, it takes $\Theta(|W|)$ for a topic. Therefore, steps 3-9 take $\Theta(|U| \times |E| \times |C| + |U| \times |E| \times |K| \times |W|)$, where $|U|$ and $|E|$ are the numbers of nodes and links, respectively. Based on the above discussions, the complexity of our algorithm is linearly related to the data size. As the amount of data increases, we can parallelize the implementation of our model.

Algorithm 1 Inference for Our Model

Require: user set u, link set E, link content W;
Ensure: user-community distribution $\boldsymbol{\pi}$, user-topic distribution $\boldsymbol{\theta}$, topic-word distribution $\boldsymbol{\phi}$, community-community distribution $\boldsymbol{\eta}$, user popularity in community $\boldsymbol{\xi}$;

```
1: Initialize α, β, ε, ρ, λ;
2: for iter = 1 : T do
3:   for each user i ∈ U do
4:     for each link e_ij ∈ E_i do
5:       Sample community c_ir by (8);
6:       Sample c'_ij by (9);
7:       Sample topic y_ir by (10);
8:     end for
9:   end for
10: end for
11: Output π, θ, φ, η, ξ by (11) - (15);
```

V. EXPERIMENTS

In this section, we evaluate our model on two datasets. We first describe the datasets used for evaluation. Second, we introduce the state-of-the-art baselines used for comparison. Third, the comparisons for the community detection results are stated. Finally, we show a case study to illustrate the ability of our model to identify the individual-level topic distribution and user popularity in communities.

A. DATASETS

To evaluate our model accurately, we use two datasets that include network topology and content. They are the Reddit dataset and the DBLP dataset [10]. Both of them supply ground truth.

The Reddit dataset is crawled from three subreddits on reddit.com: *Science*, *Movies*, and *Politics*, from August 25, 2012, to August 31, 2012. We set the community of each user according to the subreddits he or she belongs to. Therefore, the number of communities and topics are both set to 3. Reddit users are processed as nodes. A directed link with content is generated when user i replies to a post of user j.

The DBLP dataset is extracted from DBLP data. It includes 11 research conferences from 2001 to 2011. The papers belong to 3 research fields: *data mining*, *machine learning*, and *computer vision*. Therefore, the number of communities and topics are set to 3. All the authors are extracted as nodes in the network. If two authors publish a paper together, there are two directed links between them (one in each direction). The titles of the paper are processed as link content.

TABLE 2. Summary of datasets.

	# of users	# of links	#of words
Reddit	22,311	50,736	13,250
DBLP	23,422	198,224	7,457

The link content in the above two datasets is pre-processed, i.e., removing stop words and stemming. The statistics are shown in Table 2.

B. BASELINES

To validate the performance of our model for community detection, we choose four state-of-the-art baselines for comparison. The first two baselines are only based on network topology. The second two baselines use network topology and network content for community detection. They are summarized as follows.

- Louvain [15]. This is a classic community detection model based on maximizing the modularity. It does not require the number of communities as a prerequisite.
- Infomap [32]. This algorithm uses the probability flow of random walks on a network as a proxy for information flows. Then, it decomposes the network into modules by compressing a description of the probability flow.
- Community Level Diffusion (COLD) [11]. This is a generative model that generates the network topology and network content considering topic diffusion between communities. It is able to track topic changes in communities.
- Topic Correlations based Community Detection(TCCD) [30]. This is a generative model and considers topic correlations. In addition to the community structure, it can also identify the compositions of community semantics.

Our model and the baselines are all implemented on a PC with Intel 4.2GHz CPUs and 64 GB RAM.

C. METRICS

We adopt three metrics to evaluate the accuracy of the community detection results: the generalized normalized mutual information (GNMI) [33], Jaccard index, and F-score. The GNMI is widely used to evaluate the overlapping community detection performance. The Jaccard index is used to measure the similarity of two sets. Suppose that A and B are two sets. The Jaccard index is calculated by $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. The F-score is calculated by combining the precision and recall: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$

D. COMPARISONS WITH THE BASELINES

The comparisons of community detection results between our model and the baselines are shown in Table 3.

On the Reddit dataset, the Louvain and Infomap algorithms only utilize the topology information. The other two baselines and our model combine the network topology and network content information. Overall, the methods that use both types of information achieve better results than the methods that use only one type of information. Although COLD and TCCD consider community diffusion and topic correlations, our model achieves a 12.24% improvement in the GNMI, a 9.00% improvement in the Jaccard index, and a 10.14% improvement in the F-score over the second-best method: TCCD.

TABLE 3. Comparations with baselines on datasets.

	Dataset					
	Reddit			DBLP		
Method \ Metric	GNMI	Jaccard	F-score	GNMI	Jaccard	F-score
Louvain	0.18	0.42	0.51	0.29	0.59	0.73
Infomap	0.14	0.60	0.66	0.21	0.58	0.65
COLD	0.42	0.58	0.64	0.26	0.60	0.68
TCCD	0.49	0.67	0.69	0.32	0.64	0.79
CDITN(ours)	**0.55**	**0.73**	**0.76**	**0.34**	**0.66**	**0.83**

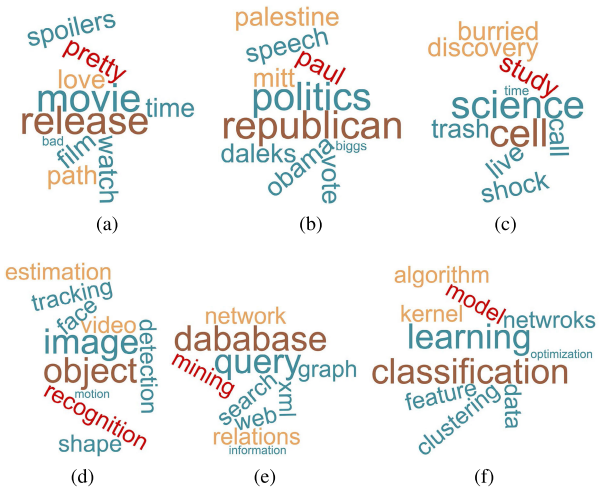


FIGURE 2. Topic-word distribution on Reddit and DBLP. (a). Topic *Movie*. (b) Topic *Politics*. (c) Topic *Science*. (d). Topic *Computer Vision*. (e) Topic *Data Mining*. (f) Topic *Machine Learning*.

On the DBLP dataset, we obtain similar results. Our model is the best according to all the metrics. The second-best method is TCCD. Our model achieves a 6.25% improvement in the GNMI, a 3.13% improvement in the Jaccard index, and a 5.06% improvement in the F-score over TCCD.

The comparisons show that considering individual topics can improve the community detection results. This method formulates the generation of the network topology and network content from the underlying idea that individual topics motivate the interactions between users and the formation of the community structure.

E. CASE STUDY

In addition to accurate community detection results, our model is capable of identifying the topic-word distribution, the user popularity with respect to topics in communities, the individual topic distribution, and the community interaction preferences, which correspond to parameters $\boldsymbol{\phi}$,$\boldsymbol{\xi}$,$\boldsymbol{\theta}$, and $\boldsymbol{\eta}$, respectively.

1) TOPIC-WORD DISTRIBUTION

Each topic is represented by a word cloud consisting of the top 10 words in the topic-word distribution. Fig. 2

TABLE 4. Top 10 users in each community in the Reddit dataset.

Community	Topic	Top 10 users
Movie	Movie	mr_majorly; CharlieDarwin2; Anomaly100; wang-banger; rottenart; JJFFMM; twolf1; davidreiss666; nowhathappenedwas; magneticsprings
	Politics	Anomaly100; mr_majorly; johnnr2; wang-banger; twolf1; JJFFMM; rottenart; CharlieDarwin2; nowhathappenedwas; davidreiss666
	Science	mr_majorly; CharlieDarwin2; Anomaly100; wang-banger; rottenart; JJFFMM; twolf1; davidreiss666; nowhathappenedwas; magneticsprings
Politics	Movie	GraybackPH; rawbamatic; maxwellhill; EthicalReasoning; johnnr2; okko7; Event0Horizon00; skcll; Citisol; BeatsBluntsAndReddit
	Politics	sdu1494; vanderlinden; Thefriendlyfaceplant; iceman2k15; steve599; AnxiousReginald; EntingFantastic; Midiex; pipboon; Primetime22
	Science	GraybackPH; rawbamatic; maxwellhill; EthicalReasoning; johnnr2; okko7; Event0Horizon00; skcll; Citisol; BeatsBluntsAndReddit
Science	Movie	Fitz11; EntingFantastic; Shogun24; Triatsila; jdCHALLENGER; Higgins420; Phaedrus78; femboost; PunchOfTheFalcon; MyJustMeHere
	Politics	Fitz11; EntingFantastic; Shogun24; Triatsila; jdCHALLENGER; Higgins420; Phaedrus78; femboost; PunchOfTheFalcon; MyJustMeHere
	Science	CharlieDarwin2; maxwellhill; twolf1; EthicalReasoning; one_of_reddit; scientologist2; Autosaver; alllie; Met4tr0n; DavidCarraway

shows that all topics identified are meaningful in both datasets.

2) USER POPULARITY WITH RESPECT TO THE TOPICS IN THE COMMUNITIES

Users in a community have different popularity levels. If a user is authoritative, his or her in-degree is larger than that of the other users. Moreover, users in a community talk about various topics; therefore, they have different popularity levels according to the different topics. Table 4 and Table 5 show the top 10 most popular users according to each topic in each community in the two datasets. It shows that users in the same community have different popularity levels with respect to different topics. In the Reddit dataset, we delete a user with an obscene user name.

3) INDIVIDUAL TOPIC DISTRIBUTIONS

In this paper, we focus on the individual-level topic distributions. Since there are a large number of users in the network, Fig. 3 illustrates the topic distribution of the top one user in each community with respect to the dominant topic in the two datasets.

4) COMMUNITY INTERACTION PREFERENCE

As discussed above, there are interactions between different communities. Fig.4 shows the probability of communication between any pair of communities. Since a community has dense inner connections, each community is more likely to interact with itself than any other community. The interactions between different communities also exist with different probabilities, which illustrates how users in different communities interact with each other.

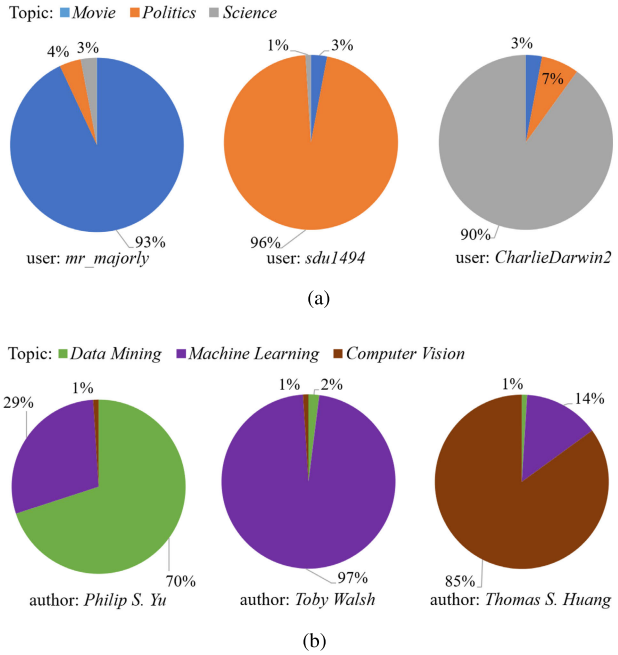


FIGURE 3. Individual topic distribution. (a). On Reddit dataset. (b) On DBLP dataset.

F. PARAMETER INITIATION

The parameters used in our model are set as follows. The number of communities and topics are set to true values according to the ground truth. All the Dirichlet hyperparameters are set to fixed values ($\rho = 0.01$, $\alpha = 0.001$, $\beta = 0.1$, and $\varepsilon = 0.001$). Our model can detect overlapping communities of each user; therefore, we set a threshold to calculate a user's communities. The threshold is set to $1/|C|$.

TABLE 5. Top 10 authors in each community in the DBLP dataset.

Community	Topic	Top 10 authors
Data Mining	Data Mining	Philip S. Yu; Yufei Tao; Nick Koudas; Surajit Chaudhuri; Jiawei Han; Xuemin Lin; Samuel Madden; Wei Wang; Wenfei Fan; Raghu Ramakrishnan
	Machine Learning	Philip S. Yu; Yufei Tao; Nick Koudas; Surajit Chaudhuri; Jiawei Han; Xuemin Lin; Samuel Madden; Wei Wang; Wenfei Fan; Raghu Ramakrishnan
	Computer Vision	Rong Jin; Jiawei Han; Zoubin Ghahramani; Wei Wang; Yong Yu; Vincent Conitzer; Nick Koudas; Wei Fan; Yee Whye Teh
Machine Learning	Data Mining	Trevor Darrell; Wenfei Fan; Wei Fan; Yihong Gong; Roberto Cipolla; Zheng Chen; Michael I. Jordan; Zhi-Hua Zhou; Yair Weiss; Vincent Conitzer
	Machine Learning	Toby Walsh; Michael I. Jordan; Vincent Conitzer; Thomas Seidl; Zoubin Ghahramani; Wei-Ying Ma; Zhi-Hua Zhou; Trevor Darrell; Spiros Papadimitriou; Yoram Singer
	Computer Vision	Zheng Chen; Toby Walsh; Rong Jin; Michael I. Jordan; Vincent Conitzer; Tao Li; Thomas Seidl; Zoubin Ghahramani; Wei-Ying Ma; Zhi-Hua Zhou
Computer Vision	Data Mining	Qiang Yang; Shuicheng Yan; Michael I. Jordan; Zhi-Hua Zhou; Xiaoou Tang; Yoram Singer; Pedro Domingos; Steven C. H. Hoi; Trevor Darrell; Yoshua Bengio
	Machine Learning	Xiaoou Tang; William T. Freeman; Trevor Darrell; Shuicheng Yan; Thomas S. Huang; Stefano Soatto; Ying Wu; Luc J. Van Gool; Yang Wang; Steven M. Seitz
	Computer Vision	Thomas S. Huang; Xiaoou Tang; William T. Freeman; Trevor Darrell; Shuicheng Yan; Stefano Soatto; Luc J. Van Gool; Steven M. Seitz; Roberto Cipolla; Serge Belongie

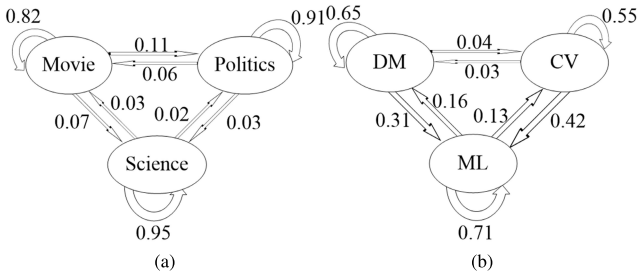


FIGURE 4. Interactions between communities. (a). Reddit dataset. (b). DBLP dataset.

If a user's probability of belonging to a community is larger than the threshold, he or she is in that community.

VI. CONCLUSION AND DISCUSSION

This paper first investigates the impact of individual-level topics on the generation of the network topology and network content. Second, we observe that individual-level topics play a core role in community generation. We find that if two individuals are in the same community and are interested in similar topics, it is more likely that a link will form between them. Third, a novel generative model for community detection is proposed by simulating the generation of the network topology and network content by considering individual topics. The mechanism revealed by our model drives the formation of links and network content and further drives the generation of communities. Finally, the experiments show that in addition to accurate community detection results, our model can identify each individual topic distribution and the most important users corresponding to different topics in each community. Investigating the evolution of communities is an important and difficult task due to the complexity of networks. In the future, we will investigate the evolution of individual topics, in particular, the impact on the evolution of communities including the community members and community semantics.

ACKNOWLEDGMENT

The authors would like to thank the two reviewers for their helpful comments and suggestions.

REFERENCES

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.
[2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
[3] M. E. J. Newman, "Communities, modules and large-scale structure in networks," *Nature Phys.*, vol. 8, no. 1, pp. 25–31, Jan. 2012.
[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] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, Feb. 2010.
[6] 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.
[7] G. Rossetti, and R. Cazabet, "Community discovery in dynamic networks: A survey," *Comput. Surv.*, vol. 51, no. 2, p. 35, Jun. 2018.
[8] N. Dakiche, F. B.-S. Tayeb, Y. Slimani, and K. Benatchba, "Tracking community evolution in social networks: A survey," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 1084–1102, May 2019.

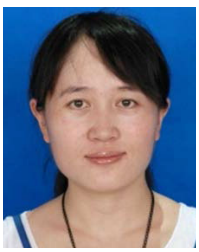
[9] D. He, Z. Feng, D. Jin, X. Wang, and W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–9.
[10] C.-D. Wang, J.-H. Lai, and P. S. Yu, "NEIWalk: Community discovery in dynamic content-based networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1734–1748, Jul. 2014.
[11] Z. Hu, J. Yao, B. Cui, and E. Xing, "Community level diffusion extraction," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1555–1569.
[12] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.
[13] R. Balasubramanyan and W. W. Cohen, "Block-LDA: Jointly modeling entity-annotated text and entity-entity links," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2011, pp. 450–461.
[14] Y. Sun, C. C. Aggarwal, and J. Han, "Relation strength-aware clustering of heterogeneous information networks with incomplete attributes," *Proc. VLDB Endowment*, vol. 5, no. 5, pp. 394–405, Jan. 2012.
[15] 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.
[16] S. E. Garza and S. E. Schaeffer, "Community detection with the label propagation algorithm: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 534, Nov. 2019, Art. no. 122058.
[17] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "A model-based approach to attributed graph clustering," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 505–516.
[18] 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.
[19] Y. Pei, N. Chakraborty, and K. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1–7.
[20] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2005, pp. 274–285.
[21] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 1089–1098.
[22] S. Pool, F. Bonchi, and M. V. Leeuwen, "Description-driven community detection," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 2, pp. 1–28, Apr. 2014.
[23] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 927–936.
[24] J. Yang and J. Leskovec, "Overlapping community detection at scale: A nonnegative matrix factorization approach," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 587–596.
[25] L. Liu, L. Xu, Z. Wangy, and E. Chen, "Community detection based on structure and content: A content propagation perspective," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 271–280.
[26] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.
[27] H. Cai, V. W. Zheng, F. Zhu, 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.
[28] G. Zhang, D. Jin, J. Gao, P. Jiao, F. Fogelman-Soulié, and X. Huang, "Finding communities with hierarchical semantics by distinguishing general and specialized topics," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3648–3654.
[29] 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.*, early access, Dec. 10, 2019, doi: 10.1109/TKDE.2019.2958806.
[30] 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.
[31] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, 2011, pp. 338–349.
[32] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008.
[33] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys.*, vol. 11, no. 3, Mar. 2009, Art. no. 033015.

HUI JIANG received the M.E. degree in engineering from the School of Computer, University of National Defense Science and Technology, China, in 2008. He is the Deputy Director of the Department of Computer Science and Technology, the Director of the Computer Experiment Teaching Center, Tianjin University Renai College, and also an Associate Professor. His main research interests include basic research on big data technology and machine learning, and application research on workflow and enterprise informatization.

LINJUAN SUN received the M.E. degree in engineering from the School of Computer Science and Software, Hebei University of Technology, Tianjin, China, in 2009. She currently works at the Department of Computer Science and Technology, Tianjin University Renai College. Her main research interests include artificial intelligence, image processing, and pattern recognition.

JUAN RAN received the M.E. degree in software engineering from Beijing University of Technology, Beijing, China. She is with the Department of Computer Science and Technology, Tianjin University Renai College. Her research interests include machine learning, data mining, and intelligence information processing.

JIANXIA BAI received the Ph.D. degree in fluid mechanics from the Department of Mechanics, School of Mechanical Engineering, Tianjin University, Tianjin, China, in 2020. She is currently a Teaching Secretary at the Department of Mathematics, Tianjin University Renai College, and also an Associate Professor. Her main research interests include basic research on experimental fluid mechanics, turbulence, and application on wavelet analysis.



XIAOYE YANG received the M.Sc. degree from the Beijing University of Aeronautics and Astronautics, in 2007. She is currently the Deputy Director and a Secretary of the Department of the Mathematics, Tianjin University Renai College, and also an Associate Professor. Her main research interests include mathematics and applied mathematics.