Integrating Group Homophily and Individual Personality of Topics Can Better Model Network Communities

Yingkui Wang
College of Intelligence and Computing
Tianjin University
Tianjin, China
ykwang@tju.edu.cn

Di Jin*
College of Intelligence and Computing
Tianjin University
Tianjin, China
jindi@tju.edu.cn

Carl Yang
Department of Computer Science
Emory University
Atlanta, USA
j.carlyang@emory.edu

Jianwu Dang
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Japan
jdang@jaist.ac.jp

***Abstract*—Community detection is an important research field in the understanding of networks. The definition of network communities focuses on denser intracommunity links and sparpser intercommunity links. It cannot explain the fundamental generation mechanisms of the two types of links, which is challenging to reveal. Unfortunately, none of existing works can solve this challenge which is important for accurately modeling community structures. This paper investigates a typical category of networks which possess contents on links. Based on analyses of real networks, we get an observation that nodes with distinctive personality regarding content topics are more active across communities, while nodes without it are more active inside a community, behaving in a similar way known as homophily. This observation provides clues to the generation of intracommunity and intercommunity links. Based on above observation, this paper proposes a novel generative community detection model called GHIPT (Group Homophily and Individual Personality of Topics) by integrating group homophily and individual personality of topics. Besides deriving more precise community results by accurately modeling intracommunity and intercommunity links, GHIPT is able to identify those nodes with distinctive personality who are more willing to interact with others from different communities. It further validates that they change their community memberships more frequently. GHIPT is evaluated on two real networks, i.e., Reddit and DBLP. Experimental results show that it outperforms all the state-of-the-art baselines. In addition to case studies on above two datasets, a case study on COVID-19 dataset provides new insights to support the ongoing fight against COVID-19 pandemic.**

***Keywords*-community detection, probabilistic graphical model, homophily, individual personality**

I. Introduction

The study of community structures in networks has been an important research topic [1], [2]. Community is defined as a group of nodes (we also call them individuals) who are densely connected inside the groups and sparsely connected across the groups [3]. Detecting accurate community structures is challenging, because links not only exist inside communities (intracommunity) but also across communities (intercommunity).

Recently, both network contents and topologies are integrated for community detection [4]–[7]. Specifically, link contents can be considered as messages transmitted among individuals, such as on Twitter, WeChat, and other online social networks.

The definition of community structure requires the best clustering of nodes with dense intracommunity links and sparse intercommunity links, which might be incorrect in some cases when considering semantics. For example, when a person in a political party frequently interacts with (e.g., cooperates or fights against) persons from other political parties, the person might be identified to be in overlapping communities incorrectly. Therefore, understanding the generation mechanisms of intracommunity and intercommunity links can promote optimal community structures from not only the perspective of topology but also semantics. Unfortunately, the issue has not been well studied by existing works. In this work, we investigate the mechanisms of link generation regarding both intracommunity and intercommunity links. Due to group homophily in networks [8], individuals in a community share similar topic interests, and they generate intracommunity links. However, based on existing research, a community not only focuses on dominant topics but also has subsidiary topics [9], [10]. Furthermore, [4] and [6] show that there are links between communities because of topic correlations. Such that, group homophily also causes intercommunity interactions.

After analyzing a large number of networks, we get a key observation that there exist a set of special individuals who

*Corresponding author

have distinctive personality regarding topics. They are more active across communities talking about various topics that are quite different from the ones shared by most of their community members. They have significant impacts on the generation of intercommunity links.

On this point, we jointly investigate the impacts of group homophily and individual distinctive personality of topics for the generation of community structures, especially their impacts on the generation of intracommunity and intercommunity links.

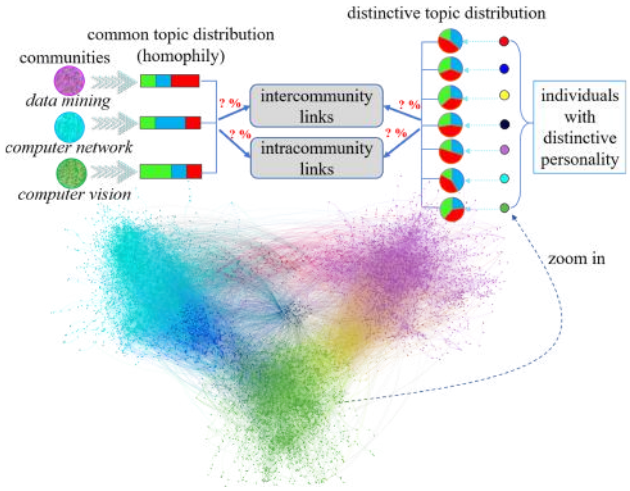


Fig. 1. The generation of intracommunity links and intercommunity links. The topology of a citation network of DBLP is shown at bottom. Purple, green and blue nodes represent community members of communities *data mining*, *computer network* and *computer vision*, respectively. Nodes with other colors are in overlapping communities. At community level (on the top left), each community possesses a common topic distribution according to group homophily. At individual level (on the top right), individuals with distinctive personality of topics are presented by small circles with colors. Our model explains the mechanisms of link generation regarding intracommunity links and intercommunity links.

Fig. 1 shows the link generation mechanism of a DBLP citation network. We extract the authors of papers and construct a directed link between two authors if one author cites the papers of another. Most of authors inside a community cite papers within the same research field and generate intracommunity links. On the other hand, when most of the authors in a research field utilize techniques from another research filed, they generate cross-disciplinary citations, i.e., intercommunity links. Moreover, we find that some authors possess distinctive personality of research topics. For example, while some authors' major research field is *data mining*, they also may have research interests in both *computer vision* and *computer networks*. They actively cite papers across communities, and play an important role in generating intercommunity links.

Some of the individuals with distinctive personality regarding topics might be in overlapping communities. Therefore, overlapping community detection is a solution to identify these individuals. However, our analyses on real networks (e.g., Reddit and DBLP) show that as much as 84.17% of the individuals with distinctive personality only belong to one community. Another issue of overlapping community detection is that it cannot identify intercommunity and intracommunity links correctly. As shown in Fig. 2, node i is in overlapping communities, i.e., *data mining* and *commuter vision*. Node j is in community *data mining*. Node i publishes a paper on topic *computer vision* and cites a paper of node j with topic *data mining*. Then, the link e_{ij} is identified as intracommunity link incorrectly. In this paper, we estimate the community indicator of source node and target node for each link, which means that two nodes of a link might be in different communities. Therefore, all links are evaluated towards whether they are inside a community or not.

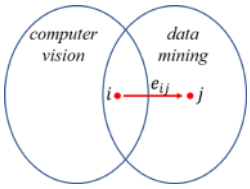


Fig. 2. An example on DBLP citation network. Community *computer vision* and community *data mining* are overlapped. e_{ij} is a directed link from node i to node j. Node i belongs to both communities. Overlapping community detection identify e_{ij} as intracommunity link incorrectly.

Based on the above observations, we consider the following three challenges.

First, how to identify individuals who are active across communities and generate intercommunity links. The challenge is important for preventing conflicts and maintaining a healthy community environment [11]. Reference [12] studies the interaction and conflict between communities in Reddit. However, to the best of our knowledge, none of existing works can actually identify the active individuals. In particular, how to model individuals who have distinctive personality of topics and are more likely to interact across communities is unknown.

Second, individual level topic distributions do not always coincide with community level. Integrating the above two aspects can improve the accuracy of community detection. Furthermore, it also improves the understanding of community semantics [13], [14]. However, how to integrate common topic distributions at community level (based on group homophily) and distinctive topic distributions at individual level (based on distinctive personality) in a seamless way is challenging and has never been studied.

Third, since networks are dynamic, individual community memberships change over time. While a large amount of existing research studies dynamic community detection [15], [16], how to capture driving factors of community evolution is still an open question. Deriving the pattern of the community evolution with regard to active individuals can provide us clues to reveal the mechanisms of community evolution. Therefore, our third challenge lies in the modeling of how individuals with distinctive personality of topics affect community evolution.

To address the above challenges, in this paper, we propose a novel probabilistic generative model called GHIPT (Group

Homophily and Individual Personality of Topics). The contributions of this work are summarized as follows:

- First, it reveals the mechanisms of link generation regarding intracommunity and intercommunity links. It for the first time captures the phenomenon that individuals with distinctive personality change their community membership more frequently.
- Second, GHIPT integrates common topic distributions at community level and distinctive topic distributions at individual level seamlessly for community detection.
- Finally, GHIPT is evaluated on two real datasets. Extensive experimental results show that it outperforms all four state-of-the-art baselines on both datasets.

II. Related Work

Community detection. Earlier studies mainly focus on network topology to detect community structure by its definition [3], [17]–[19]. As network content provides valuable information to node attributes or link semantics, it implies underlying reasons of community formation. For example, nodes with similar attributes are more likely to be in the same community. A large number of community detection models have been proposed by integrating network topology and network content [2], [20], [21]. Some of them both use node content and link content [6], [22]. While others only use link content to investigate the mechanisms of link generation and further the mechanisms of community generation. In addition to accurate community detection results, network content also makes the understanding of community semantics available [14], [23]. Community profiling is proposed by [5]. The work of [6] investigates topic correlations in community structure and explains community semantics in a natural way. Many recent studies leverage graph neural networks for joint node embedding and community detection [24]–[27]. Our method identifies whether a link is inside a community or across communities by using link content to achieve accurate community structure.

Interaction between communities. In social networks, the interaction reflects social opinion propagation. Recent works analyze interactions between communities [4], [11], [12], [28]–[30]. The work of [12] investigates the generation process of conflicts that occur from one community to another. The interactions between communities are significant for maintaining network environment. Intercommunity interaction and conflict in Reddit are first studied by [12]. It reveals the mechanisms of the interactions between communities. Reference [4] studies community level diffusion in social networks. References [31] and [32] study community conflict. It is important for preventing conflicts and maintaining a healthy community environment [11]. Therefore, detecting individuals who are active across communities and are more likely to initiate community interaction is a key issue, which is to be resolved by this paper.

III. The Model

A. Problem Formulation

We first describe our problem formulation. The notations used in this work are summarized in Table I.

TABLE I
Notations

Notations	Descriptions
U, E, W	Set of users, links, and link contents
K, C, V	Set of topics, communities, and vocabulary
e_{ij}	Directed link from node i to node j
$c^i_{e_{ij}}$	Source-node community indicator specific to e_{ij}
$g^j_{e_{ij}}$	Target-node community indicator specific to e_{ij}
W_{ij}, W_{ijq}	Word list of e_{ij}, and the q-th word of W_{ij}
$\boldsymbol{\pi}_i$	Multinomial distribution over communities specific to user i
$\boldsymbol{\theta}_c$	Multinomial distribution over topics specific to community c
$\boldsymbol{\chi}_i$	Multinomial distribution over topics specific to user i
$\boldsymbol{\tau}_i$	Bernoulli distribution over homophily and distinctive personality specific to user i
$\boldsymbol{\phi}_k$	Multinomial distribution over words specific to topic k
$\boldsymbol{\eta}_{ck}$	Multinomial distribution over communities specific to community c talking about topic k
$\boldsymbol{\xi}_c$	Multinomial distribution over all users specific to community c
k_{ij}	Topic indicator of link e_{ij}
s_{ij}	The indicator of where the topic of link e_{ij} is from. If $s_{ij} = 1$, link topic is from individual topic distribution. If $s_{ij} = 0$, link topic is from community topic distribution.
$\boldsymbol{\sigma}$, λ, δ, α, ε, ρ, β	Parameters of Dirichlet priors

Definition 1. A **network** G comprises of user set U, edge set E, and edge content set W, i.e., $G = (U, E, W)$. A directed link from node i to node j is denoted by e_{ij}. The edge content of node i's outgoing edge e_{ij} is denoted by W_{ij}.

Definition 2. At community level, the **content of a community** c is a multinomial distribution $\boldsymbol{\theta}_c$ over topics. $\boldsymbol{\theta}_{ck}$ denotes the probability that the topic of a link is talking about k when the source node is in community c.

Definition 3. At individual level, the **individual content** is a multinomial distribution $\boldsymbol{\chi}_i$ over topics. $\boldsymbol{\chi}_{ik}$ denotes the probability that individual i is interested in topic k.

Definition 4. Individual i's **characteristic** is defined by a Bernoulli distribution $\boldsymbol{\tau}_i$. It represents the probability that the topic of link e_{ij} is decided by homophily or distinctive personality of individual i when i starts a link.

Definition 5. Individual i's **community membership** is a multinomial distribution $\boldsymbol{\pi}_i$ over communities. $\boldsymbol{\pi}_{ic}$ denotes the probability of belonging to community c for i.

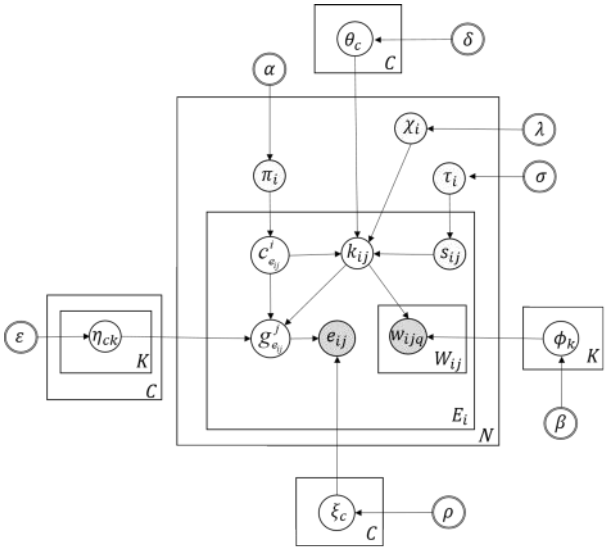


Fig. 3. The graphical representation of GHIPT.

Definition 6. **Community preference** of communities corresponding to a topic is defined by a multinomial distribution $\boldsymbol{\eta}_{ck}$ over communities. $\boldsymbol{\eta}_{ck,g}$ denotes the probability of interacting with individuals in community g for individuals in community c when they talk about topic k.

Definition 7. **Individual popularity** in a community g is a multinomial distribution $\boldsymbol{\xi}_g$ over all individuals. Each dimension $\boldsymbol{\xi}_{gj}$ denotes the probability that node i is selected as target node in community g.

Definition 8. A **topic** is a multinomial distribution $\boldsymbol{\phi}_k$ over vocabularies. $\boldsymbol{\phi}_{kq}$ denotes the probability of belonging to topic k for word q.

B. Model Structure

We propose a probabilistic generative model with two components, i.e., topology generation and link content generation. Fig. 3 shows the probabilistic graphical representation of this model.

Topology generation component. Consider the generation of a directed link e_{ij}. e_{ij} is either inside a community or across two different communities. The source-node community is sampled from $\boldsymbol{\pi}_i$, i.e., $c^i_{e_{ij}}$. Next, the key issue is how the topic of e_{ij} is sampled, i.e., k_{ij}.

If i has distinctive personality, the topic might be different from its community's topic preference. Otherwise, homophily plays a dominant role in deciding topic k and this topic is more likely consistent with its community's topic preference. We sample a switch s_{ij} from $\boldsymbol{\tau}_i$. If s_{ij} is equal to 1, k_{ij} is sampled from i's individual topic distribution $\boldsymbol{\chi}_i$. If s_{ij} is equal to 0, it is from i's community topic distribution $\boldsymbol{\theta}_{c^i_{e_{ij}}}$.

Then, we evaluate where the target node j is from, i.e., community $g^j_{e_{ij}}$. We highlight that $g^j_{e_{ij}}$ is not sampled from node j's community distribution $\boldsymbol{\pi}_j$. Instead, it is sampled based on $\boldsymbol{\eta}_{ck}$ ($c = c^i_{e_{ij}}, k = k_{ij}$). Finally, we sample individual j from $\boldsymbol{\xi}_{g^j_{e_{ij}}}$.

Link Content Generation Component. In the topology generation component, we already get the topic of each link, i.e., k_{ij}. Each word in the link content is sampled from $\phi_{k_{ij}}$. Following ideas of LDA [33], all words on links are generated.

Generative process. The generative process is summarized as follows.

1) For each community c in C
   a) Sample its topic distribution from a Dirichlet prior: $\boldsymbol{\theta}_c \mid \boldsymbol{\delta} \sim Dir(\boldsymbol{\delta})$;
   b) Sample its user distribution from a Dirichlet prior: $\boldsymbol{\xi}_c \mid \rho \sim Dir(\rho)$;
   c) For each topic k in K
      i) Sample community distribution for community c and topic k from a Dirichlet prior: $\boldsymbol{\eta}_{ck} \mid \varepsilon \sim Dir(\varepsilon)$;
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 user i in U
   a) Sample its community distribution from a Dirichlet prior: $\boldsymbol{\pi}_i \mid \alpha \sim Dir(\alpha)$;
   b) Sample individual topic distribution from a Dirichlet prior: $\boldsymbol{\chi}_i \mid \lambda \sim Dir(\lambda)$;
   c) Sample personality distribution from a Beta prior: $\boldsymbol{\tau}_i \mid \boldsymbol{\sigma} \sim Beta(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$;
   d) For each directed link e_{ij} in E_i
      i) Sample source-node community indicator $c^i_{e_{ij}}$ from a Multinomial distribution: $c^i_{e_{ij}} \mid \boldsymbol{\pi}_i \sim Mul(\boldsymbol{\pi}_i)$;
      ii) Sample indicator s_{ij} from a Bernoulli distribution: $s_{ij} \mid \boldsymbol{\tau}_i \sim Ber(\boldsymbol{\tau}_i)$;
      iii) If $s_{ij} = 0$, sample topic indicator k_{ij} from a Multinomial distribution: $k_{ij} \mid \boldsymbol{\theta}_{c^i_{e_{ij}}} \sim Mul(\boldsymbol{\theta}_{c^i_{e_{ij}}})$. If $s_{ij} = 1$, sample k_{ij} from a Multinomial distribution: $k_{ij} \mid \boldsymbol{\chi}_i \sim Mul(\boldsymbol{\chi}_i)$;
      iv) Sample target-node community indicator $g^j_{e_{ij}}$ from a Multinomial distribution: $g^j_{e_{ij}} \mid \boldsymbol{\eta}_{c^i_{e_{ij}},k_{ij}} \sim Mul(\boldsymbol{\eta}_{c^i_{e_{ij}},k_{ij}})$;
      v) Sample target node j of link e_{ij} from a Multinomial distribution: $e_{ij} \mid \boldsymbol{\xi}_{g^j_{e_{ij}}} \sim Mul(\boldsymbol{\xi}_{g^j_{e_{ij}}})$
      vi) For each word w_{ijq} in W_{ij}
         A) Sample word from a Multinomial distribution: $w_{ijq} \mid \boldsymbol{\phi}_{k_{ij}} \sim Mul(\boldsymbol{\phi}_{k_{ij}})$;

C. Model Inference

Based on the probabilistic graphical model, the posterior distribution of GHIPT is shown by Eq. (1). U, E, and W are observed data. s, k, c, and g are latent variables. Set $H = \{\boldsymbol{\sigma}, \lambda, \delta, \alpha, \varepsilon, \rho, \beta\}$ includes all hyper parameters. Our target is to infer parameters $\{\boldsymbol{\tau}, \boldsymbol{\chi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\phi}\}$ by optimizing (1).

$$
\begin{aligned}
&P(\boldsymbol{\tau}, \boldsymbol{\chi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\phi}, s, k, c, g | U, E, W, H) \\
&\propto P(\boldsymbol{\tau}|\boldsymbol{\sigma}) P(s|\boldsymbol{\tau}) P(\boldsymbol{\chi}|\lambda) P(\boldsymbol{\theta}|\delta) P(c|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\alpha) \\
&\cdot P(g|\boldsymbol{\eta}, c, k) P(\boldsymbol{\eta}|\varepsilon) P(k|\boldsymbol{\theta}, \boldsymbol{\chi}, c, s) P(e|\boldsymbol{\xi}, g) P(\boldsymbol{\xi}|\rho) \\
&\cdot P(\boldsymbol{\phi}|\beta) P(w|\boldsymbol{\phi}, k).
\end{aligned} \tag{1}
$$

In Eq. (1), we find that it is hard to calculate the normalizing constant. Therefore, we adopt Collapsed Gibbs Sampling [34] for approximate inference.

The first step is to marginalize out all parameters, i.e., $\{\boldsymbol{\tau}, \boldsymbol{\chi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\phi}\}$. We get Eq. (2).

$$
\begin{aligned}
&P(s, k, c, g|.) \\
&\propto \int P(\boldsymbol{\pi}|\alpha) P(c|\boldsymbol{\pi}) d\boldsymbol{\pi} \\
&\cdot \int P(\boldsymbol{\tau}|\boldsymbol{\sigma}) P(s|\boldsymbol{\tau}) d\boldsymbol{\tau} \\
&\cdot \int P(\boldsymbol{\chi}|\lambda) P(k|\boldsymbol{\chi}, s=1) d\boldsymbol{\chi} \\
&\cdot \int P(\boldsymbol{\theta}|\delta) P(k|\boldsymbol{\theta}, c, s=0) d\boldsymbol{\theta} \\
&\cdot \int P(\boldsymbol{\eta}|\varepsilon) P(g|\boldsymbol{\eta}, c, k) d\boldsymbol{\eta} \\
&\cdot \int P(\boldsymbol{\xi}|\rho) P(e|\boldsymbol{\xi}, g) d\boldsymbol{\xi} \\
&\cdot \int P(\boldsymbol{\phi}|\beta) P(w|\boldsymbol{\phi}, k) d\boldsymbol{\phi}.
\end{aligned} \tag{2}
$$

The first integral in Eq. (2) is calculated as follows.

$$
\begin{aligned}
&\int P(\boldsymbol{\pi}|\alpha) P(c|\boldsymbol{\pi}) d\boldsymbol{\pi} \\
&= \prod_{i=1}^{|U|} \frac{\Gamma(|C|\alpha_i)}{(\Gamma(\alpha_i))^{|C|}} \cdot \frac{\prod_{c=1}^{|C|} \Gamma(n_i^{(c)} + \alpha_i)}{\Gamma(n_i^{(\cdot)} + |C|\alpha_i)},
\end{aligned} \tag{3}
$$

where n_i^c is the number of links of user i that are assigned to community c. Dots in all equations denote marginal counts. $n_i^{(\cdot)}$ is the total number of links that are assigned to all communities for user i.

The second integral in Eq. (2) is calculated as follows.

$$
\begin{aligned}
&\int P(\boldsymbol{\tau}|\boldsymbol{\sigma}) P(s|\boldsymbol{\tau}) d\boldsymbol{\tau} \\
&= \prod_{i=1}^{|U|} \left(\frac{1}{B(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)}\right) \prod_{j=1}^{|E_i|} B(s_{ij} + \boldsymbol{\sigma}_1, 1 - s_{ij} + \boldsymbol{\sigma}_2),
\end{aligned} \tag{4}
$$

where $P(\boldsymbol{\tau}|\boldsymbol{\sigma})$ follows Beta distribution. $\boldsymbol{\sigma}_1$ corresponds to $s_{ij} = 1$, which means that the topic is from distinctive topic distribution. $\boldsymbol{\sigma}_2$ corresponds to $s_{ij} = 0$, which means that the topic is from common topic distribution. $P(s|\boldsymbol{\tau})$ follows Bernoulli distribution. $B(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ is the Beta function.

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

$$
\begin{aligned}
&\int P(\boldsymbol{\chi}|\lambda) P(k|\boldsymbol{\chi}, s=1) d\boldsymbol{\chi} \\
&= \prod_{i=1}^{|U|} \frac{\Gamma(|K|\lambda)}{(\Gamma(\lambda))^{|K|}} \cdot \frac{\prod_{k=1}^{|K|} \Gamma(n_i^{(k)} + \lambda)}{\Gamma(n_i^{(\cdot)} + |K|\lambda)},
\end{aligned} \tag{5}
$$

where $n_i^{(k)}$ is the number of links of user i that are assigned to topic k. $n_i^{(\cdot)}$ is total number of links that are assigned to all topics for user i.

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

$$
\begin{aligned}
&\int P(\boldsymbol{\theta}|\delta) P(k|\boldsymbol{\theta}, c, s=0) d\boldsymbol{\theta} \\
&= \prod_{c=1}^{|C|} \frac{\Gamma(|K|\delta)}{(\Gamma(\delta))^{|K|}} \cdot \prod_{i=1}^{|U|} \frac{\prod_{k=1}^{|K|} \Gamma(n_i^{(ck)} + \delta)}{\Gamma(n_i^{(c\cdot)} + |K|\delta)},
\end{aligned} \tag{6}
$$

where $n_i^{(ck)}$ is the number of user i's links assigned to community c specific to topic k. $n_i^{(c\cdot)}$ is the total number of user i's links aggregating all topics specific to community c.

The fifth integral in Eq. (2) is calculated by Eq. (7)

$$
\begin{aligned}
&\int P(\boldsymbol{\eta}|\varepsilon) P(g|\boldsymbol{\eta}, c, k) d\boldsymbol{\eta} \\
&= \prod_{c=1}^{|C|} \prod_{k=1}^{|K|} \frac{\Gamma(|C|\varepsilon)}{(\Gamma(\varepsilon))^{|C|}} \cdot \frac{\prod_{m=1}^{C} \Gamma(n_E^{(ck,m)} + \varepsilon)}{\Gamma(n_E^{(ck,\cdot)} + |C|\varepsilon)},
\end{aligned} \tag{7}
$$

where $n_E^{(ck,m)}$ is the number of links whose target nodes are assigned to community m and source nodes are in community c talking about topic k. $n_E^{(ck,\cdot)}$ is the number of all links whose target nodes are assigned to all communities for source nodes in community c and talking about topic k.

The sixth integral in Eq. (2) is calculated by Eq. (8)

$$
\begin{aligned}
&\int P(\boldsymbol{\xi}|\rho) P(e|\boldsymbol{\xi}, g) d\boldsymbol{\xi} \\
&= \prod_{g=1}^{|C|} \frac{\Gamma(|U|\rho)}{(\Gamma(\rho))^{|U|}} \cdot \frac{\prod_{u=1}^{|U|} \Gamma(n_E^{(gu)} + \rho)}{\Gamma(n_E^{(g\cdot)} + |U|\rho)},
\end{aligned} \tag{8}
$$

where $n_E^{(gu)}$ is the number of times that user u is selected as target node from community g for all links in a network. $n_E^{(g\cdot)}$ is the marginal counts over all users in community g.

The seventh integral in Eq. (2) is calculated by Eq. (9)

$$
\begin{aligned}
&\int P(\boldsymbol{\phi}|\beta) P(w|\boldsymbol{\phi}, k) d\boldsymbol{\phi} \\
&= \prod_{k=1}^{|K|} \frac{\Gamma(|V|\beta)}{(\Gamma(\beta))^{|V|}} \cdot \frac{\prod_{w=1}^{|V|} \Gamma(n_E^{(kw)} + \beta)}{\Gamma(n_E^{(k\cdot)} + |V|\beta)},
\end{aligned} \tag{9}
$$

where $n_E^{(k,w)}$ is the number of times that word w is assigned to topic k for all links in a network. $n_E^{(k,\cdot)}$ is the number of times that all words are assigned to topic k for all links in the network.

The second step is to sample all latent variables. For each link e_{ij}, we sample user i's community membership $c_{e_{ij}}^i$.

$$
\begin{aligned}
&P(c_{e_{ij}}^i = c | c_{\neg ij}, k_{ij} = k, s_{ij} = 0, g = m, .) \\
&= \frac{P(c, k, s, g|.)}{P(c_{\neg ij}, k, s, g|.)} \\
&= \frac{n_{i,\neg ij}^{(ck)} + \delta}{n_{i,\neg ij}^{(c\cdot)} + |K|\delta} \cdot \frac{n_{i,\neg ij}^{(c)} + \alpha}{n_{i,\neg ij}^{(\cdot)} + |C|\alpha} \frac{n_{E,\neg ij}^{(ck,m)} + \varepsilon}{n_{E,\neg ij}^{(ck,\cdot)} + |C|\varepsilon},
\end{aligned} \tag{10}
$$

where $\neg ij$ means excluding link e_{ij}. The indicator s_{ij} is sampled by following equations.

$$P(s_{ij}=s|s_{\neg ij},c_{ij}=c,k_{ij}=k,g=m,.) = \Psi(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2)\cdot\frac{n^{(k)}_{i,\neg ij}+\lambda}{n^{(\cdot)}_{i,\neg ij}+|K|\lambda}\cdot\frac{n^{(ck)}_{i,\neg ij}+\delta}{n^{(c\cdot)}_{i,\neg ij}+|K|\delta}, \tag{11}$$

$$\Psi(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2)=\begin{cases}B(1+\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2) & s_{ij}==1\\ B(\boldsymbol{\sigma}_1,1+\boldsymbol{\sigma}_2) & s_{ij}==0.\end{cases} \tag{12}$$

The topic of each link is sampled as follows.

$$P(k_{ij}=k|k_{\neg ij},c_{ij}=c,s_{ij}=s,g=m,.) = \omega(s)\cdot\frac{n^{(ck,m)}_{E,\neg ij}+\varepsilon}{n^{(ck,\cdot)}_{E,\neg ij}+|C|\varepsilon} \cdot\frac{\prod_{w=1}^{|V|}\prod_{q=0}^{n^{(w)}_{ij}-1}(n^{(w)}_{k,\neg ij}+q+\beta)}{\prod_{q=0}^{n^{(\cdot)}_{ij}-1}(n^{(\cdot)}_{k,\neg ij}+q+\beta)}, \tag{13}$$

$$\omega(s)=\begin{cases}\dfrac{n^{(k)}_{i,\neg ij}+\lambda}{n^{(\cdot)}_{i,\neg ij}+|K|\lambda} & s==1\\ \dfrac{n^{(ck)}_{i,\neg ij}+\delta}{n^{(c\cdot)}_{i,\neg ij}+|K|\delta} & s==0,\end{cases} \tag{14}$$

where $n^{(w)}_{ij}$ is the number of times that word w appears in link content W_{ij}.

For the target user j, its community $g^j_{e_{ij}}$ is sampled as follows.

$$P(g^j_{e_{ij}}=c|g^j_{e_{ij},\neg ij},k_{ij}=k,s_{ij}=s,c=m,.) = \frac{n^{(ck,m)}_{E,\neg ij}+\varepsilon}{n^{(ck,\cdot)}_{E,\neg ij}+|C|\varepsilon}\cdot\frac{n^{(gu)}_{E,\neg ij}+\rho}{n^{(g\cdot)}_{E,\neg ij}+|U|\rho}. \tag{15}$$

D. Parameter estimation

Parameters $\hat{\boldsymbol{\pi}}_{ic}$ and $\hat{\boldsymbol{\tau}}_i$ are estimated by following equations:

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

$$\hat{\boldsymbol{\tau}}_i=\frac{n^{(1)}_i+\boldsymbol{\sigma}_1}{n^{(\cdot)}_i+\boldsymbol{\sigma}_1+\boldsymbol{\sigma}_2}. \tag{17}$$

Parameters $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\chi}}$, $\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{\xi}}$, and $\hat{\boldsymbol{\phi}}$ are estimated according to δ, λ, ε, ρ, and β similarly.

E. Time Complexity Analysis

The algorithm of GHIPT is illustrated in Alg. 1. The numbers of topics and communities are fixed to $|K|$ and $|C|$ respectively. T denotes the number of iterations for convergence. For each link of a user, step 5 samples community indicator of source node. Equation (10) takes a constant time, because all counters are stored in memory. The calculation of steps 6 and 8 all take a constant time. At step 7, equation (13) takes $\Theta(|V|)$ for a topic. Therefore, steps 5-8 take $\Theta(|U|\times|E|\times|C|+|U|\times|E|\times|K|\times|V|)$, where $|U|$ and $|E|$ are the number of nodes and the number of links. In a summary, the complexity of GHIPT is nearly linearly related to data size.

Algorithm 1 Inference of the GHIPT model

Input: user set U, edge set E, edge content W;

Output: user-community distribution $\boldsymbol{\pi}$, user-topic distribution $\boldsymbol{\chi}$, community-topic distribution $\boldsymbol{\theta}$, topic-word distribution $\boldsymbol{\phi}$, community preference of communities corresponding to a topic $\boldsymbol{\eta}$, user popularity in community $\boldsymbol{\xi}$, individual characteristic $\boldsymbol{\tau}$;

1: Initialize $\alpha,\beta,\varepsilon,\rho,\lambda,\sigma,\delta$;
2: **for** $iter=1:T$ **do**
3: **for** each user $i\in U$ **do**
4: **for** each link $e_{ij}\in E_i$ **do**
5: Sample community indicator of source node $c^i_{e_{ij}}$ via (10);
6: Sample indicator s_{ij} via (11);
7: Sample topic indicator k_{ij} via (13);
8: Sample community indicator of target node $g^j_{e_{ij}}$ via (15);
9: **end for**
10: **end for**
11: **end for**
12: Output $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{\chi}$, $\boldsymbol{\tau}$;

IV. Experiments

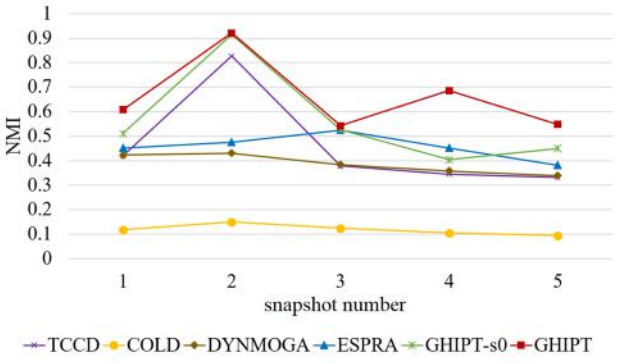
To evaluate the performance of GHIPT, we choose two real datasets, i.e., a social network Reddit [22] and a citation network DBLP [35]. Both datasets are supplied with ground truth. Reddit dataset is extracted from four sub-forums, i.e., *movie*, *science*, *politics* and *olympics*. It is divided into five snapshots, which includes 46,594 users and 21,130, 18,809, 20,085, 23,317, and 32,019 links at the five snapshots respectively. For the DBLP citation network, we collect papers in three research fields, i.e., *data mining*, *computer vision* and *computer network* from 2013 to 2018 with each year as one snapshot. We extract each paper's first and last authors as nodes and construct citation relations. If author i publishes a paper that cites a paper of author j, a directed link e_{ij} is generated with author i's paper title as link content. It includes 21,542 authors and consists of 15,631, 72,895, 156,347, 249,343, 297,371, and 129,324 links at the six snapshots respectively.

GHIPT is compared with four state-of-the-art baselines: i) TCCD [6], a generative model considering topic correlations in social networks; ii) COLD [4], a generative model for identifying temporal topics of communities; iii) ESPRA [36], an evolutionary clustering algorithm combining structural

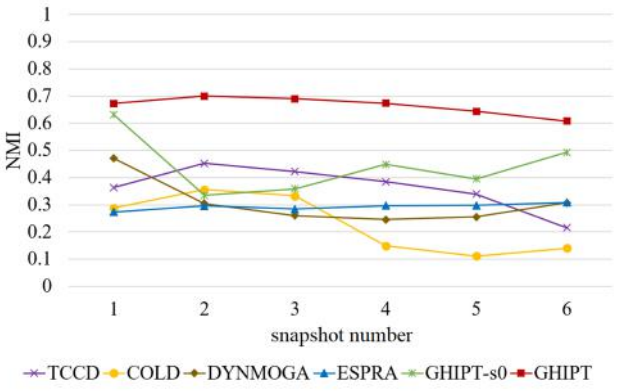
perturbation and topological features; and iv) DYNMOGA [37], a multi-objective approach to detect communities. To validate that GHIPT can observe individuals with distinctive personality, we make a variation of GHIPT by setting $s = 0$ denoted by GHIPT-s0, in which all link topics are derived from community topic distributions while ignoring distinctive individual topic distributions. Setting $s = 1$ is also implemented, but we get no results because of the huge amount of parameters. So, we ignore the baseline with $s = 1$.

Parameters of all baselines are set as suggested by their authors. In GHIPT, the values of hyper parameters are set as follows: $\boldsymbol{\sigma}_1 = 1$, $\boldsymbol{\sigma}_2 = 100$, $\lambda = 0.01$, $\delta = 0.001$, $\alpha = 0.01$, $\varepsilon = 0.1$, $\rho = 0.001$, and $\beta = 0.1$. For the numbers of communities and topics, we set them to values according to ground-truth.

We adopt GNMI (Generalized Normalized Mutual Information) [38] and F-score as metrics.



(a)



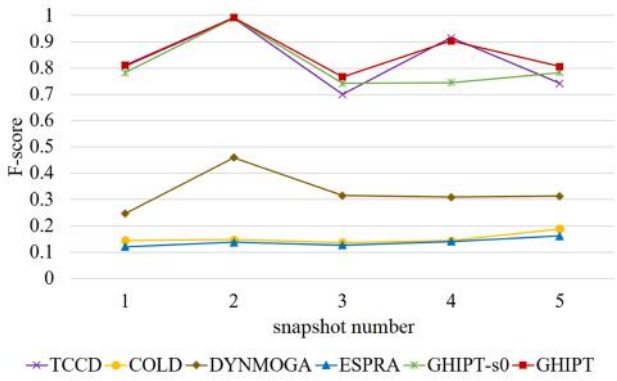
(b)

Fig. 4. Comparisons of community detection results with respect to NMI. (a) is on the network of Reddit, and (b) is on the network of DBLP.

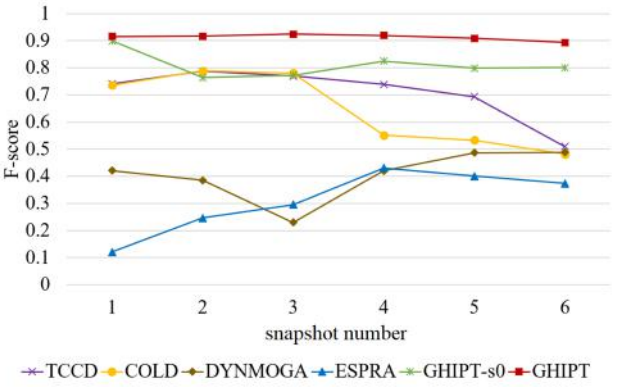
A. Results Comparison

Fig. 4 and Fig. 5 demonstrate the comparisons of community detection results on the two datasets. On Reddit dataset, Fig. 4(a) shows that GHIPT outperforms all baselines at all snapshots in terms of NMI metric. GHIPT and GHIPT-s0 are the best and the second best methods respectively at snapshots 1, 2, and 5. GHIPT improves 19.03%, 0.76%, 3.34%, 52%, and 21.79% compared with the second best methods at each snapshot. The results show that integrating group homophily and distinctive personality of topics is efficient for community detection. The comparisons between GHIPT and GHIPT-s0 indicate that considering distinctive individual topic distributions is significant. For the F-score metric, Fig. 5(a) shows that GHIPT are the best methods at snapshots 1, 2, 3, and 5. It improves 0.62%, 0.32%, 3.38% and 8.67% of the second best methods.

On the citation network of DBLP, Fig. 4(b) shows that GHIPT outperforms all baselines at all snapshots in terms of NMI metric. GHIPT and GHIPT-s0 are the best and the second best methods at snapshots 1, 4, 5, and 6. GHIPT improves 6.59%, 54.61%, 63.2%, 50.35%, 62.73%, and 23.25% compared with second best methods at each snapshot. The comparisons between GHIPT and GHIPT-s0 also confirm the effectiveness of considering distinctive individual topic distributions. For F-score metric, Fig. 5(b) shows that GHIPT are the best methods at all snapshots. It improves 0.62%, 1.84%, 16.13%, 18.51%, 11.42%, 13.84%, and 11.64% of the second best methods.



(a)



(b)

Fig. 5. Comparisons of community detection results with respect to F-score. (a) is on the network of Reddit, and (b) is on the network of DBLP.

B. Case Studies

Recall the first and the third challenges. First, we illustrate the identified individuals with distinctive personality and an-

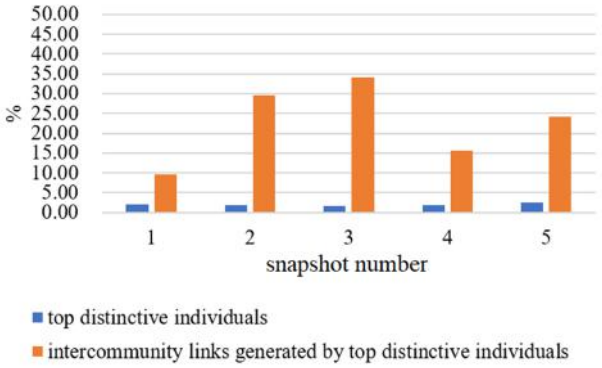


Fig. 6. Distinctive individuals identified on Reddit.

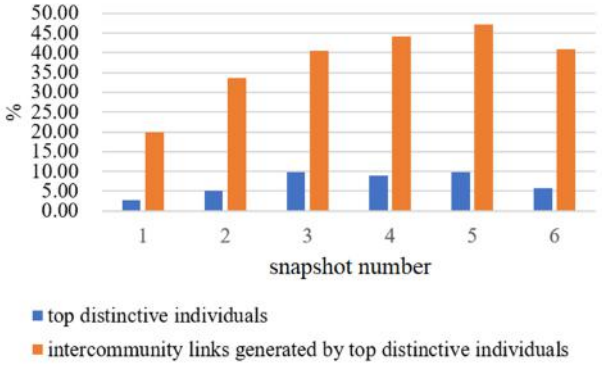


Fig. 7. Distinctive individuals identified on DBLP.

alyze the contributions they made to intercommunity links. Second, we illustrate their community evolution. Datasets used in these studies include Reddit, DBLP as well as the COVID-19 Open Research Dataset Challenge (CORD-19)[1] corpus (202003-13). The first two datasets supply ground-truth, therefore, we first illustrate case studies on them and then illustrate a case study on CORD-19 without ground-truth.

1) Case Studies on Reddit and DBLP: For the value of parameter τ, we set a threshold to 0.01 to obtain distinctive individuals. For an individual, if τ_i is larger than 0.01, he or she is identified as a distinctive individual. Fig. 6 and Fig. 7 show the distinctive individuals we found on two datasets. On Reddit, distinctive individuals account for 2.05%, 1.95%, 1.74%, 1.9%, and 2.59% of their community members. They generate 9.64%, 29.52%, 34.15%, 15.54%, and 24.13% intercommunity links. On the citation network DBLP, distinctive individuals account for 2.76%, 5%, 9.73%, 8.94%, 9.82%, and 5.74% of their community members. They generate 19.98%, 33.73%, 40.47%, 44.22%, 47.17%, and 40.87% intercommunity links. The results show that a small number of individuals with distinctive personality generate a large number of intercommunity links.

Fig. 8 and Fig. 9 illustrate community evolution of all individuals and distinctive individuals on these two datasets. They show the transfer of community members from one snapshot (y-axes) to next snapshot (x-axes). On Reddit, it is difficult to observe the transfer pattern at snapshot 1 and 2 because of the changing number of communities. Fig. 8(a) shows the transfer from snapshot 3 to snapshot 4. Most of individuals in community C_1 and community C_2 at snapshot 3 remain in their communities at snapshot 4. Individuals in community C_3 transfer to community C_2 and C_3 partly. By comparison, the first figure in Fig. 8(b) shows that most of distinctive individuals in community C_2 transfer to community C_3.

On DBLP, Fig. 9 shows that the distinctive individuals are more likely to change their community memberships at all snapshots. Therefore, if a community includes too many distinctive individuals, its members will also change frequently; and vice versa.

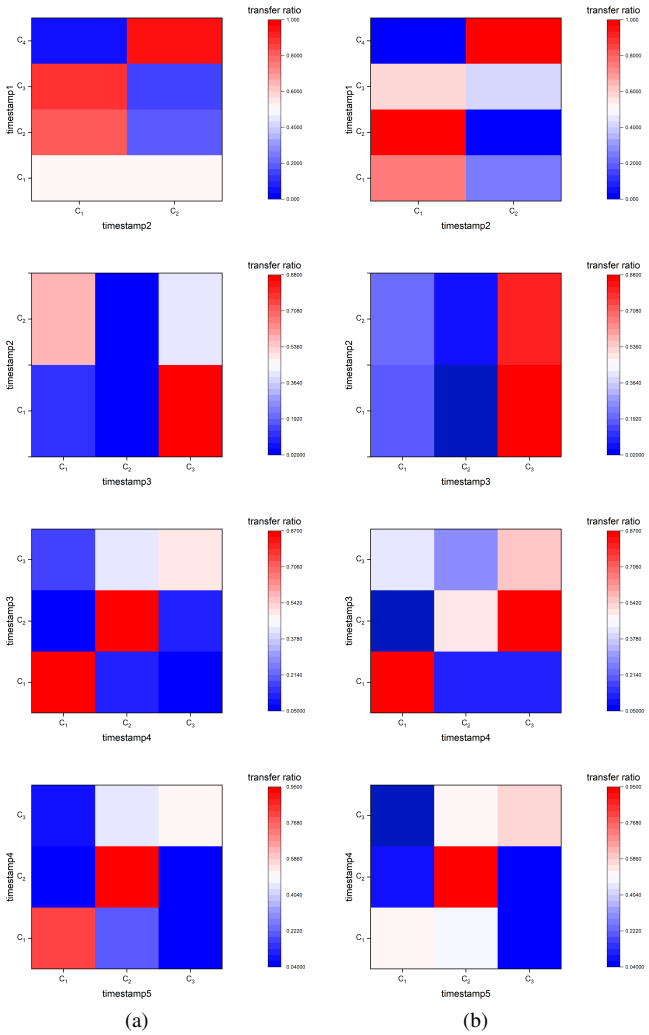


Fig. 8. Community evolution on Reddit. Column (a) considers all individuals, and column (b) considers individuals with distinctive personality only.

2) A Case Study on CORD-19: As coronavirus disease 2019 (COVID-19) spreads globally, on March 16th, 2020,

[1]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

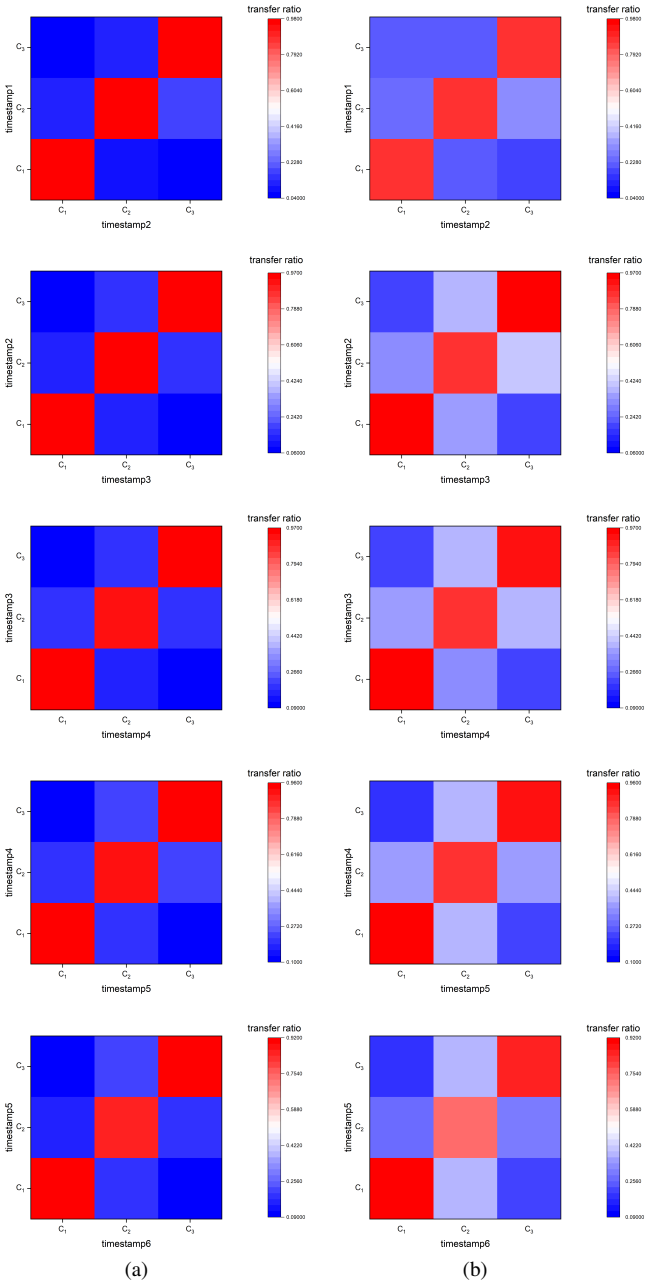


Fig. 9. Community evolution on citation network. Column (a) considers all individuals, and column (b) considers individuals with distinctive personality only.

the White House and a coalition of leading research groups released the COVID-19 Open Research Dataset (CORD-19) that consists of over 141,000 scholarly articles, about COVID-19, SARS-CoV-2, and related coronaviruses. COVID-19 is processed into an author citation network like DBLP network. It includes 215,349 authors who publish more than 4 papers in the dataset. It is divided into two snapshots. The first snapshot includes papers published before December 1st, 2019 when the first case of coronavirus disease was found. Other papers are in the second snapshot. There are 61,987 and 466,607 links in snapshots 1 and 2, respectively. The number of communities and topics are all set to 20 according to [39].

Topics SARS and CORD-19. In GHIPT, ϕ_k is a multinomial distribution over words specific to topic k. We focus on two topics, i.e., *SARS* and *CORD-19* in the COVID-19 dataset. They are represented by word clouds consisting of the top 30 words in each topic. Fig. 10(a) and Fig. 10(b) show that the two topics identified are meaningful in two snapshots, respectively. We can conclude that researchers of COVID-19 mainly focus on subjects of *covid*, *transmissibles*, *infections*, *globally*, etc., which is urgent to defeat the new virus.

Fig. 10. Topic-word distribution. (a) is the topic *SARS*, and (b) is the topic *CORD-19*.

Top 10 Most Cited Authors in Community SARS. Table II shows the top 10 most cited authors in community *SARS*. We validate manually that all of the authors are famous experts in the field of "Viruses".

TABLE II
TOP 10 MOST CITED AUTHORS IN COMMUNITY *SARS*.

Community	**Authors**
SARS	Wesley I.Sundquist; Patrick CY Woo; Ron A M Fouchier; Christian Drosten; Jasper Fuk-Woo Chan ; Vincent J Munster; Sean K. Lau; Jiyong Zhou; Xavier de Lamballerie; Gregory B Melikyan

Authors with Cross-disciplinary Researches. In GHIPT, the value of parameter τ indicates if an individual is distinctive and more active across communities, which correctly corresponds to cross-disciplinary researchers in citation network. Table III shows 10 out of 637 authors with cross-disciplinary researches in snapshot 2. We validate manually that these authors' research fields include "Bioinformatics", "Cell", "Mathematical epidemiology", "Viruses", "Statistics", "Data Sciences", etc. To defeat the new virus, the above cross-disciplinary researchers can provide critical understanding of the virus besides the research filed of "Viruses".

V. CONCLUSION AND DISCUSSIONS

This paper investigates the impacts of group homophily and individual distinctive personality on community detection. It essentially interprets the mechanisms of intracommunity and intercommunity link generation. The experimental results on two real datasets show that GHIPT is able to resolve

TABLE III
10 AUTHORS WITH CROSS-DISCIPLINARY RESEARCHES.

Snapshot	Authors
2	Yanni Sun; Gail Rosen; Gerardo Chowell; Jacco Wallinga; Samuel Alizon; Qi Wang; Hongjie Yu; Weiwei Guo; Lauren Ancel Meyers; Bruno Coutard

three challenges: (1) It identifies individuals with distinctive personality who are more active across communities and generate intercommunity links; (2) It is a novel unified generative model integrating group homophily and individual distinctive personality and achieves state-of-the-art community detection results; (3) It for the first time explains the phenomenon that individuals with distinctive personality change their community membership more frequently. However, the changing pattern of individual characteristics over time is not investigated in this work. It leads individuals to participate in different communities regarding topics dynamically, which will be investigated in our future work.

ACKNOWLEDGMENT

The research was supported by the National Natural Science Foundation of China (No. 61772361). We would like to thank Dr. Pengfei Jiao for providing the useful advices for this work.

REFERENCES

[1] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[2] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *TKDD*, vol. 8, no. 1, pp. 1–28, 2014.

[3] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, no. 12, pp. 7821–7826, 2002.

[4] Z. Hu, J. Yao, B. Cui, and E. Xing, "Community level diffusion extraction," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1555–1569.

[5] H. Cai, V. W. Zheng, F. Zhu, K. C.-C. Chang, and Z. Huang, "From community detection to community profiling," *Proceedings of the VLDB Endowment*, vol. 10, no. 7, pp. 817–828, 2017.

[6] Y. Wang, D. Jin, K. Musial, and J. Dang, "Community Detection in Social Networks Considering Topic Correlations," *AAAI*, vol. 33, pp. 321–328, 2019.

[7] L. Yang, F. Wu, J. Gu, C. Wang, X. Cao, D. Jin, and Y. Guo, "Graph attention topic modeling network," in *WWW*, 2020, pp. 144–154.

[8] C. R. Shalizi and A. C. Thomas, "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies," *Sociological Methods & Research*, vol. 40, no. 2, pp. 211–239, 2010.

[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 *AAAI*, 2017.

[10] D. Jin, X. Wang, R. He, D. He, J. Dang, and W. Zhang, "Robust detection of link communities in large social networks by exploiting link semantics." in *AAAI*, 2018.

[11] W. L. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec, "Loyalty in Online Communities." *AAAI*, vol. 2017, pp. 540–543, 2017.

[12] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community Interaction and Conflict on the Web," *WWW*, pp. 933–943, 2018.

[13] 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," *AAAI*, pp. 5730–5736, 2019.

[14] 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 *IJCAI*, 2018, pp. 3648–3654.

[15] F. Liu, J. Wu, S. Xue, C. Zhou, J. Yang, and Q. Sheng, "Detecting the evolving community structure in dynamic social networks," *WWW*, pp. 1–19, 2019.

[16] P. Jiao, W. Yu, W. Wang, X. Li, and Y. Sun, "Exploring temporal community structure and constant evolutionary pattern hiding in dynamic networks," *Neurocomputing*, vol. 314, pp. 224–233, 2018.

[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[18] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 26113, 2004.

[19] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *WWW*, 2008, pp. 695–704.

[20] 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," *TKDE*, 2019, DOI: 10.1109/TKDE.2019.2937298.

[21] D. Jin, X. Wang, D. He, J. Dang, and W. Zhang, "Robust detection of link communities with summary description in social networks," *TKDE*, 2019, DOI: 10.1109/TKDE.2019.2958806.

[22] C.-D. Wang, J.-H. Lai, and S. Y. Philip, "Neiwalk: community discovery in dynamic content-based networks," *TKDE*, vol. 26, no. 7, pp. 1734–1748, 2014.

[23] D. He, Y. Song, and D. Jin, "A simple and effective community detection method combining network topology with node attributes," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2019, pp. 168–175.

[24] L. Yang, Y. Guo, J. Gu, D. Jin, B. Yang, and X. Cao, "Probabilistic graph convolutional network via topology-constrained latent space model," *IEEE Transactions on Cybernetics*, pp. 1–14, 2020.

[25] L. Yang, F. Wu, Y. Wang, J. Gu, and Y. Guo, "Masked graph convolutional network," in *IJCAI 2019, Macao, China, August 10-16*, 2019, pp. 4070–4077.

[26] C. Yang, M. Liu, Z. Wang, L. Liu, and J. Han, "Graph clustering with dynamic embedding," *arXiv preprint arXiv:1712.08249*, 2017.

[27] C. Yang, H. Lu, and K. C.-C. Chang, "Cone: Community oriented network embedding," *arXiv preprint arXiv:1709.01554*, 2017.

[28] C. Hauser, "Reddit bans nazi groups and others in crackdown on violent content," *New York Times*, 2017.

[29] V. Belák, S. Lam, and C. Hayes, "Cross-community influence in discussion fora," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[30] H. Tajfel, J. C. Turner, W. G. Austin, and S. Worchel, "An integrative theory of intergroup conflict," *Organizational identity: A reader*, vol. 56, p. 65, 1979.

[31] S. Chen, X. Cai, B. Li, and Z. Hou, "Community conflict prediction method based on spliced bilstm," in *IVPAI*, vol. 11321. International Society for Optics and Photonics, 2019, p. 113211V.

[32] Y. Chen and X. Ye, "Online community conflict decomposition with pseudo spatial permutation," in *International Conference on Computational Data and Social Networks*. Springer, 2019, pp. 246–255.

[33] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[34] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[35] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *KDD*. ACM, 2008, pp. 990–998.

[36] P. Wang, L. Gao, and X. Ma, "Dynamic community detection based on network structural perturbation and topological similarity," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, no. 1, p. 013401, 2017.

[37] F. Folino and C. Pizzuti, "An evolutionary multiobjective approach for community discovery in dynamic networks," *TKDE*, vol. 26, no. 8, pp. 1838–1852, 2013.

[38] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *KDD*. ACM, 2009, pp. 877–886.

[39] W. Xuan, S. Xiangchen, L. Bangzheng, G. Yingjun, and H. Jiawei, "Comprehensive named entity recognition on cord-19 with distant or weak supervision," in *2020 Intelligent Systems for Molecular Biology (ISMB'20)*, 2020.