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

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

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

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 sparser 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...
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.

Therefore, all links are evaluated towards whether they are inside a community or not.

![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.](image)

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-based Topic Model).
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 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>
<thead>
<tr>
<th>Notations</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U, E, W$</td>
<td>Set of users, links, and link contents</td>
</tr>
<tr>
<td>$K, C, V$</td>
<td>Set of topics, communities, and vocabulary</td>
</tr>
<tr>
<td>$c_{ij}$</td>
<td>Source-node community indicator specific to $e_{ij}$</td>
</tr>
<tr>
<td>$g_{ij}$</td>
<td>Target-node community indicator specific to $e_{ij}$</td>
</tr>
<tr>
<td>$W_{ij}, W_{ijq}$</td>
<td>Word list of $e_{ij}$, and the $q$-th word of $W_{ij}$</td>
</tr>
<tr>
<td>$\pi_i$</td>
<td>Multinomial distribution over communities specific to user $i$</td>
</tr>
<tr>
<td>$\theta_c$</td>
<td>Multinomial distribution over topics specific to community $c$</td>
</tr>
<tr>
<td>$\chi_i$</td>
<td>Multinomial distribution over topics specific to user $i$</td>
</tr>
<tr>
<td>$\tau_i$</td>
<td>Bernoulli distribution over homophily and distinctive personality specific to user $i$</td>
</tr>
<tr>
<td>$\phi_k$</td>
<td>Multinomial distribution over words specific to topic $k$</td>
</tr>
<tr>
<td>$\eta_{ck}$</td>
<td>Multinomial distribution over communities specific to community $c$ talking about topic $k$</td>
</tr>
<tr>
<td>$\xi_c$</td>
<td>Multinomial distribution over all users specific to community $c$</td>
</tr>
<tr>
<td>$k_{ij}$</td>
<td>Topic indicator of link $e_{ij}$</td>
</tr>
<tr>
<td>$s_{ij}$</td>
<td>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.</td>
</tr>
<tr>
<td>$\sigma, \lambda, \delta$</td>
<td>Parameters of Dirichlet priors</td>
</tr>
<tr>
<td>$\alpha, \varepsilon, \rho, \beta$</td>
<td>Parameters of Dirichlet priors</td>
</tr>
</tbody>
</table>

**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 $\theta_c$ over topics. $\theta_{ck}$ 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 $\chi_i$ over topics. $\chi_{ik}$ denotes the probability that individual $i$ is interested in topic $k$.

**Definition 4.** Individual $i$’s characteristic is defined by a Bernoulli distribution $\tau_i$. 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 $\pi_i$ over communities. $\pi_{ic}$ denotes the probability of belonging to community $c$ for $i$. 

Definitions of individual attributes and community memberships.
Definition 6. **Community preference** of communities corresponding to a topic is defined by a multinomial distribution \( \eta_{ck} \) over communities. \( \eta_{ck,g} \) 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 \( \xi_{gij} \) over all individuals. Each dimension \( \xi_{gij} \) denotes the probability that node \( i \) is selected as target node in community \( g \).

Definition 8. A **topic** is a multinomial distribution \( \phi_{kq} \) over vocabularies. \( \phi_{kq} \) 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 \( \pi_i \), i.e., \( c^i_{sij} \). 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 \( \tau_i \). If \( s_{ij} \) is equal to 1, \( k_{ij} \) is sampled from \( i \)'s individual topic distribution \( \chi_i \). If \( s_{ij} \) is equal to 0, it is from \( i \)'s community topic distribution \( \theta_{c_{sij}} \).

Then we evaluate where the target node \( j \) is from, i.e., community \( g^j_{tij} \). We highlight that \( g^j_{tij} \) is not sampled from node \( j \)'s community distribution \( \pi_j \). Instead, it is sampled based on \( \eta_{ck} (c = c^j_{tij}, k = k_{ij}) \). Finally, we sample individual \( j \) from \( \xi_{g^j_{tij}} \).

**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: \( \theta_c \mid \delta \sim Dir(\delta) \);
   b) Sample its user distribution from a Dirichlet prior: \( \xi_c \mid \rho \sim Dir(\rho) \);
   c) For each topic \( k \) in \( K \)
      i) Sample community distribution for community \( c \) and topic \( k \) from a Dirichlet prior: \( \eta_{c,k} \mid \varepsilon \sim Dir(\varepsilon) \);

2) For each topic \( k \) in \( K \)
   a) Sample word distribution from a Dirichlet prior: \( \phi_k \mid \beta \sim Dir(\beta) \);

3) For each user \( i \) in \( U \)
   a) Sample its community distribution from a Dirichlet prior: \( \pi_i \mid \alpha \sim Dir(\alpha) \);
   b) Sample individual topic distribution from a Dirichlet prior: \( \chi_i \mid \lambda \sim Dir(\lambda) \);
   c) Sample personality distribution from a Beta prior: \( \tau_i \mid \sigma \sim Beta(\sigma_1, \sigma_2) \);
   d) For each directed link \( e_{ij} \) in \( E_i \)
      i) Sample source-node community indicator \( c^i_{sij} \) from a Multinomial distribution: \( c^i_{sij} \mid \pi_i \sim Mul(\pi_i) \);
      ii) Sample indicator \( s_{ij} \) from a Bernoulli distribution: \( s_{ij} \mid \tau_i \sim Ber(\tau_i) \);
      iii) If \( s_{ij} = 0 \), sample topic indicator \( k_{ij} \) from a Multinomial distribution: \( k_{ij} \mid \theta_{c_{sij}} \sim Mul(\theta_{c_{sij}}) \).
      iv) If \( s_{ij} = 1 \), sample \( k_{ij} \) from a Multinomial distribution: \( k_{ij} \mid \chi_i \sim Mul(\chi_i) \);
      v) Sample target-node community indicator \( g^j_{tij} \) from a Multinomial distribution: \( g^j_{tij} \mid \eta_{g^j_{tij}} \sim Mul(\eta_{g^j_{tij}}) \);
      vi) Sample target node \( j \) of link \( e_{ij} \) from a Multinomial distribution: \( e_{ij} \mid \phi_{g^j_{tij}} \sim Mul(\phi_{g^j_{tij}}) \)

   A) Sample word from a Multinomial distribution: \( w_{ij} \mid \phi_{k_{ij}} \sim Mul(\phi_{k_{ij}}) \);

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 \( H = \{\sigma, \lambda, \alpha, \varepsilon, \rho, \beta\} \) includes all hyper parameters. Our target is to infer parameters \( \{\tau, \chi, \theta, \pi, \eta, \xi, \phi\} \) by optimizing (1).
\[
P(\tau, \chi, \theta, \pi, \eta, \xi, \phi, s, k, c, g|U, E, W, H) \\
\propto P(\tau|\sigma)P(s|\tau)P(\chi|\lambda)P(\theta|\pi)P(c|\pi)P(\pi|\alpha) \\
\cdot P(g|\eta, c, k)P(\eta|\xi)P(s|\theta, \chi, c, s)P(e|\xi, g)P(\xi|\rho) \\
\cdot P(\phi|\beta)P(w|\phi, k).
\]

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

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

\[
\int P(\pi|\alpha)P(c|\pi)d\pi \\
\int P(\tau|\sigma)P(s|\tau)d\tau \\
\int P(\chi|\lambda)P(k|\chi, s = 1)d\chi \\
\int P(\theta|\delta)P(\theta|\pi, c, s = 0)d\theta \\
\int P(\eta|\xi)P(\eta|\xi, c, k)d\eta \\
\int P(\xi|\rho)P(e|\xi, g)d\xi \\
\int P(\phi|\beta)P(w|\phi, k)d\phi.
\]

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

\[
\int P(\pi|\alpha)P(c|\pi)d\pi \\
= \prod_{i=1}^{[U]} \frac{\Gamma([C]\alpha_i)}{\Gamma(\alpha_i)^{[C]}} \cdot \frac{\prod_{i=1}^{[C]} \Gamma(n_i^{(c)} + \alpha_i)}{\Gamma(n_i^{(c)} + [C] \alpha_i)} ,
\]

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. \(n_i^{(c)}\) 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.

\[
\int P(\tau|\sigma)P(s|\tau)d\tau \\
= \prod_{i=1}^{[U]} \left( \frac{1}{B(\sigma_1, \sigma_2)} \right) \prod_{j=1}^{[E]} \cdot B(s_{ij} + \sigma_1, 1 - s_{ij} + \sigma_2),
\]

where \(P(\tau|\sigma)\) follows Beta distribution. \(\sigma_1\) corresponds to \(s_{ij} = 1\), which means that the topic is from distinctive topic distribution. \(\sigma_2\) corresponds to \(s_{ij} = 0\), which means that the topic is from common topic distribution. \(P(s|\tau)\) follows Bernoulli distribution. \(B(\sigma_1, \sigma_2)\) is the Beta function.

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

\[
\int P(\chi|\lambda)P(k|\chi, s = 1)d\chi \\
= \prod_{i=1}^{[U]} \frac{\Gamma([K]\lambda)}{\Gamma(\lambda)^{[K]}} \cdot \frac{\Gamma(n_i^{(k)} + \lambda)}{\Gamma(n_i^{(k)} + [K] \lambda)},
\]

where \(n_i^{(k)}\) is the number of links of user \(i\) that are assigned to topic \(k\). \(n_i^{(c)}\) 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)

\[
\int P(\theta|\delta)P(\theta|\pi, c, s = 0)d\theta \\
= \prod_{k=1}^{[K]} \frac{\Gamma([K]\delta)}{\Gamma(\delta)^{[K]}} \cdot \prod_{i=1}^{[U]} \frac{\Gamma(n_i^{(ck)} + \delta)}{\Gamma(n_i^{(c)} + [K] \delta)},
\]

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

\[
\int P(\eta|\xi)P(\eta|\xi, c)d\eta \\
= \prod_{c=1}^{[C]} \frac{\Gamma([C]|\xi)}{\Gamma(\xi)^{[C]}} \cdot \prod_{g=1}^{[G]} \frac{\Gamma(n_{(ck,m)}^{(g)}) + \varepsilon}{\Gamma(n_{(c,k)}^{(g)} + [C] \varepsilon)},
\]

where \(n_{(ck,m)}^{(g)}\) is the number of links whose target nodes are assigned to community \(m\) and source nodes are in community \(c\) talking about topic \(k\). \(n_{(c,k)}^{(g)}\) 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)

\[
\int P(\xi|\rho)P(e|\xi, g)d\xi \\
= \prod_{g=1}^{[G]} \frac{\Gamma([U]|\rho)}{\Gamma(\rho)^{[U]}} \cdot \frac{\Gamma(n_{(g,w)}^{(u)}) + \rho}{\Gamma(n_{(w)}^{(g)} + [U] \rho)},
\]

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

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

\[
\int P(\phi|\beta)P(w|\phi, k)d\phi \\
= \prod_{k=1}^{[K]} \frac{\Gamma([V]|\beta)}{\Gamma(\beta)^{[V]}} \cdot \frac{\Gamma(n_{(k,w)}^{(u)}) + \beta}{\Gamma(n_{(w)}^{(k)} + [V] \beta)},
\]

where \(n_{(k,w)}^{(u)}\) is the number of times that word \(w\) is assigned to topic \(k\) for all links in a network. \(n_{(w)}^{(c)}\) 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 \(c_{e_{ij}}\).

\[
P(c_{e_{ij}} = c|c_{-e_{ij}}, k_{ij} = k, s_{ij} = 0, g = m, .) \\
= \frac{P(e_{-e_{ij}}, k_{ij}, s_{ij} = 0, g = m, .)}{P(e_{-e_{ij}}, k_{ij}, s_{ij} = 0, g = m, .)} \\
= \frac{n_{i_{-s_{ij}}}^{(ck)} + \delta}{n_{i_{-s_{ij}}}^{(c)} + [K] \delta} \cdot \frac{n_{s_{ij}}^{(c)} + \alpha}{n_{s_{ij}}^{(c)} + \alpha n_{E_{-s_{ij}}}^{(c,k)} + \varepsilon},
\]

where \(-s_{ij}\) means excluding link \(e_{ij}\). The indicator \(s_{ij}\) is sampled by following equations.
The topic of each link is sampled as follows.

\[ P(s_{ij} = s|s_{-ij}, c_{ij} = c, k_{ij} = k, g = m_{ij}) = \Psi(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \cdot \frac{n_{(k)}_{i_{-ij}}^{(c)} + \lambda}{n_{(i_{-ij})}^{(c)} + |K|\lambda}, \frac{n_{(c)}_{i_{-ij}}^{(k)} + \delta}{n_{i_{-ij}}^{(k)} + |K|\delta}. \] (11)

\[ \Psi(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = \begin{cases} B(1 + \sigma_1, \sigma_2) & s_{ij} = 1 \\ B(\sigma_1, 1 + \sigma_2) & s_{ij} = 0. \end{cases} \] (12)

The topic of each link is sampled as follows.

\[ P(k_{ij} = k|k_{-ij}, c_{ij} = c, s_{ij} = s, g = m_{ij}) = \omega(s) \cdot \frac{n_{(c_k,m)}^{(ck)} + \varepsilon}{n_{E_{-ij}}^{(ck)} + |C|\varepsilon}, \] (13)

\[ \omega(s) = \begin{cases} \frac{n_{(i_{-ij})}^{(c)} + \lambda}{n_{i_{-ij}}^{(c)} + |K|\lambda} & s = 1 \\ \frac{n_{(i_{-ij})}^{(c)} + \delta}{n_{i_{-ij}}^{(c)} + |K|\delta} & s = 0. \end{cases} \] (14)

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

For the target user \( j \), its community \( g_{c_{ij}} \) is sampled as follows.

\[ P(g_{c_{ij}} = c|g_{-c_{ij}}, k_{ij} = k, s_{ij} = s, c = m_{ij}) = \frac{n_{E_{-ij}}^{(ck)} + \rho}{n_{E_{-ij}}^{(ck)} + |U|\rho}. \] (15)

D. Parameter estimation

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

\[ \hat{\pi}_{ic} = \frac{n_{(i)}^{(c)} + \alpha}{n_{i}^{(c)} + |C|\alpha}. \] (16)

\[ \hat{\tau}_i = \frac{n_{(i)}^{(c)} + \sigma_1}{n_{i}^{(c)} + \sigma_1 + \sigma_2}. \] (17)

Parameters \( \hat{\theta}, \hat{\chi}, \hat{\eta}, \hat{\xi}, \phi, \Psi \) are estimated according to \( \delta, \lambda, \varepsilon, \rho, \) and \( \beta \) 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.

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: $\sigma_1 = 1$, $\sigma_2 = 100$, $\lambda = 0.01$, $\delta = 0.001$, $\alpha = 0.01$, $\epsilon = 0.1$, $\rho = 0.001$, 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.

![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.](image1)

![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.](image2)

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.

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.

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.

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

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.

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

Authors with Cross-disciplinary Researches. In GHIPT, the value of parameter $\tau$ 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
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