Community detection across multiple social networks based on overlapping users

Ziqing Zhu | Tao Zhou | Chenghao Jia | Weijia Liu | Bo Liu | Jiuxin Cao

1School of Computer Science and Engineering, Southeast University, Jiangsu, China
2School of Cyber Science and Engineering, Southeast University, Jiangsu, China

Correspondence
Jiuxin Cao, Jiangsu Provincial Key Laboratory of Computer Networking Technology, Southeast University, Nanjing, Jiangsu, 211189, China. Email:jx.cao@seu.edu.cn

Funding information
Jiangsu Provincial Key Laboratory of Computer Networking Technology, Jiangsu Provincial Key Laboratory of Network and Information Security, Grant/Award Number: BM2003201; Jiangsu Provincial Key Project, Grant/Award Number: BE2018706; Key Laboratory of Computer Network and Information Integration of Ministry of Education of China, Grant/Award Number: 93K-9; National Natural Science Foundation of China, Grant/Award Numbers: 61772133, 61972087; National Social Science Foundation of China, Grant/Award Number: 19@ZH014; Natural Science Foundation of Jiangsu Province, Grant/Award Number: SBK2019022870

Abstract
With the rapid development of Internet technology, online social networks have got fast development and become increasingly popular. Meanwhile, the research works across multiple social networks attract growing attention, among which community detection is quite important for online security problems, such as the user behavior analysis and abnormal community discovery. In this paper, a community detection method is proposed across multiple social networks based on overlapping users. First, the concept of overlapping users is defined, then an algorithm CMN_NMF is designed to discover the stub communities from overlapping users. After that, we expand each stub community in different social networks by adding the users with strong similarity, and in the end different communities are derived out across networks. At last, experiments are carried out to verify the reasonability of community. Our method shows better performance in real datasets compared with others.

1 INTRODUCTION

In recent years, with the rapid development of Internet technology and the large-scale popularization of smart mobile terminals, online social networks (OSNs) have gradually become important platforms for people's daily communication and information sharing. However, the easy accessibility and wide connectivity of OSNs make it possible for those malicious users to achieve ulterior purpose, such as rumor spreading, spam comments, information theft, and so on. It brings the severe security challenge to OSN services. At the same time, social network, as the online projection of human social relations in reality, contain the big value of data. Many works take advantage of big data in OSNs to analyze public opinions, monitor emergency, and trace social interaction, etc. Among them, community detection work is an important research.
Social networks generally have cluster feature which is similar to the real society. This cluster feature can be defined as the concept of community. By analyzing the potential pattern behind the community, we can explore the information diffusion model, and discover the evolution rules in social network structure. The community reflects the correlation between individuals in the social network. The study of the community plays a crucial role in understanding the network structure, information spread, and other characteristics of networks.

Relevant studies\(^1\) show that due to the differences in the functions of social networks, single social network is difficult to meet diverse demands of users, so users generally have multiple accounts across social networks. Such users are called overlapping users across multiple social networks. The general existence of overlapping users enriches data of single social network and is important in information diffusion across social networks. Community detection based on overlapping users makes it possible for many research works across OSNs such as behavior analysis and information diffusion.

Community across multiple social network is based on overlapping users, who build the bridge to integrate information from different social networks. Compared with the works in single network, community detection across multiple platforms is different mainly in the following three aspects: (1) It is based on overlapping users who connect different platforms; (2) Multidimensional information from heterogeneous OSNs is fused for detection and the community structures of individual network are refined by overlapping users (single network’s community detection does not require this process, since there are no such special users as overlapping users); (3) It emphasizes the diffusion of information from different networks rather than different types of multimodal data, and it also requires the discovery of information diffusion pattern by information spreading path, in which three kinds of user relations are extracted, that is, “normal-overlapping,” “overlapping-overlapping,” and “normal-normal.”

In this study, we propose a cross-social network community detection method based on overlapping users. First, the overlapping users are extracted from multiple social networks, and then are divided into multiple stub communities with a novel community detection algorithm. Users in same stub community are closely related with each other. After that, each stub community expands in different social networks by merging new users. Finally, the cross-social network communities are extracted once each stub community ceases growing. As the example illustrated in Figure 1, users A, B, C, D, E, and F are overlapping users who have different accounts on three networks: OSN1, OSN2, and OSN3. These users can compose a heterogeneous integrated network. Then, users are divided into different stub communities based on this integrated network, and the cross-network communities can be obtained by expanding stub communities in three networks: OSN1, OSN2, and OSN3. The user group in solid cycle in Figure 1 represents a stub community of the integrated network.

The remainder of the paper is organized as follows. Section 2 reviews the related works. Section 3 gives the concept of overlapping users and the overlapping user network model. The community detection method is analyzed in Section 4. Section 5 evaluates the experiments. Finally, the last section concludes this paper’s work and discusses the future research work.

**Figure 1** Community across multiple social networks
With the rapid development of big data technology and complex network research, community detection has become a popular topic in academic world.

In earlier research, many scholars pay attention to the study of community structure based on graph topology. Girvan et al.\(^2\) proposed the method of community detection using edge betweenness. The concept of modularity proposed by Newman et al.\(^3\) becomes an important standard for evaluating community detection performance. Then, Newman,\(^4\) Clauset,\(^5\) and Blondel et al.\(^6\) proposed a variety of community detection algorithms specific to optimizing modularity. Considering that one member of a community can belong to different communities at the same time in the real world, Palla et al.\(^7,8\) defined a complete subgraph structure with K vertices, called K-clique. Community structure is formed based on K-clique. Furthermore, community detection algorithm based on seed diffusion was proposed by Lancichinetti et al.\(^9\) which took the top-N different users with the highest influence in the network as seed nodes, the community detection is realized by calculating the closeness between ordinary nodes and seed nodes. Wang et al.\(^10\) applied nonnegative matrix factorization (NMF) to the work of community detection, and proposed community partition method on directed weighted network and undirected weighted network, respectively.

In OSNs, there are many types of relations between users, and using only 1D attribute data could fail to obtain the ideal community detection. Many works focus on diverse data. Zhao et al.\(^11\) quantified the user's textual content using TF-IDF in the Twitter network, together with the user's “follow” relation to detect the community by k-means clustering. Xu et al.\(^12\) built complex generation model based on the latent Dirichlet allocation (LDA) by integrating users' temporal, spatial, social, and other type of relations, to solve and obtain the probability distribution of users in the location-based communities. Liu et al.\(^13\) proposed a multi-NMF algorithm which can decompose multidimensional data between nodes in the network to detect community. Through adding regularization term of multi-NMF algorithm, He et al.\(^14\) proposed two kinds of CoNMF community detection algorithms based on pair-wise and cluster-wise approaches. Based on the Twitter network, Pei et al.\(^15\) constructed a three-layer relation matrix by integrating user-user relation, user-message relation, and message-message relation, and used joint nonnegative matrix three-factor decomposition algorithm to obtain user membership in communities.

Furthermore, beyond focusing on single social network, a lot of research works are carried out for multiple social networks. Based on the idea of same users have similar clustering in different networks, Nguyen et al.\(^16\) fused different nodes in multiple networks into a unified model, using asymmetric nonnegative matrix decomposition algorithm to integrate connections for all network, and obtained community. Yu et al.\(^17\) studied the differences of community relations among overlapping users across social networks, and used spectral clustering algorithm to find community structure based on the confidence and discrepancy features of relations between users.

It is not difficult to find through the above introduction that the network community detection using matrix decomposition approach has become relatively mainstream technology. Meanwhile, related works do not fully explore the diversity of the association between overlapping users in multiple social networks, ignoring the limitations of matrix decomposition in the application of large social networks. In view of the above problems, this paper provides a more suitable method for cross-social network community detection.

### 3 | CONCEPT AND MODEL

Some related concept definitions and the integrated network model will be first described here to better explain our research work; the details are as follows:

**Definition 1** (Multiple social networks). This paper models multiple social networks as a set \( G \) of graphs. \( G \) consists of \( p \) single social networks. Each social network is expressed as \( G^i = (U^i, E^i) \), where \( U^i \) is the set of users and \( E^i \) is set of users' relations in the \( i \)-th network.

**Definition 2** (Overlapping users). One user can use different social networks. This paper defines users who have multiple accounts on different social networks as overlapping users.

**Definition 3** (Integrated network model). Considering overlapping users are crucial in community for information diffusion across social networks, this paper models a integrated network based on overlapping users. The integrated network is denoted by \( G^0 = (U^0, E^0) \), where \( U^0 \) is the set of all cross-social network overlapping users, and \( E^0 \) is the set
of all overlapping users’ fused association relations, including “follow,” “comment,” “like,” and “semantic similarity” relations. Based on the diversity of correlations, the overlapping user integrated network $G^0$ is divided into different subnetworks $G_x^0 = (U^0, \phi^x)$, $\phi^x$ is edge set on “follow,” “comment,” “like,” and “semantic similarity” relation edge, respectively.

**Definition 4** (Stub community). This paper defines the communities which consist of overlapping users on integrated network as stub communities, which are core basis of communities across multiple social networks.

Based on the above description, our research work of community detection across multiple social networks focuses on the solution of the following two problems.

**Problem 1.** Stub community detection based on integrated network. Given integrated network $G^0 = (U^0, E^0)$, the task is to find the community membership for each overlapping user.

**Problem 2.** Community expansion in each single social network. Given each social network $G^i = (U^i, E^i)$, stub community set $N^0$, the task is to add new users to these stub communities, each stub community will naturally grow up into a community across multiple social networks.

Compared with other works for community detection across multiple social networks, our research has the following advantages:

1. A integrated network model is specially designed for overlapping users to integrate the multiple dimensional information from different social networks, which benefits the community detection.
2. The stub community origins from overlapping user integrated network, and it is the same as the source of information diffusion across multiple social networks.
3. The community detection is divided into two phases, from the stub communities on overlapping users to communities across social networks, so the method is scalable for large social networks.

## 4 | CROSS-SOCIAL NETWORK COMMUNITY DETECTION

### 4.1 | Stub community detection algorithm

Overlapping users in integrated networks present multidimensional association relations. In order to more effectively use these multidimensional relations, a **Cross Multi-Network** community detection algorithm CMN_NMF based on **NMF** is proposed deliberately. NMF is a data dimensionality reduction method, which has been widely used in image analysis, text clustering, and other fields, which has explainable, effective, and flexible features. Here it is used for stub community detection.

#### 4.1.1 | User adjacency matrix construction

Four types of weighted user adjacency matrices are constructed, corresponding to the “follow,” “comment,” “like,” and “semantic similarity” relations in the overlapping user integrated network. Elements in the first three type adjacency matrices are the number of “follow,” “comment,” “like” relations between users respectively. These asymmetric adjacency matrices represent weighted networks formed by directed relations. The elements in the last type of matrix describe the semantic similarity between users of the topic distribution based on their post content, and this matrix is symmetric to indicate weighted networks formed by undirected relations.

#### 4.1.2 | Objective function

This paper proposes a stub community detection algorithm CMN_NMF for overlapping users by multidimensional integrated network nonnegative matrix decomposition.

In CMN_NMF algorithm, the stub community membership matrix can be obtained by decomposing user adjacency matrices. It can be seen from Equation (1) that the factorization for a asymmetric matrix is $U^{(0)}H^{(0)}(U^{(0)})^T$, and for a
symmetric matrix is $W^{(g)}(W^{(g)})^T$, here a common community membership matrix $S$ is used as constraint condition to reach global optimization, and the matrix $S$ is the stub community detection we want.

$$\begin{align*}
\min_{U^{(i)}, W^{(g)}, S} & \sum_{t=1}^{p} a_t||A^{(t)} - U^{(i)}H^{(i)}(U^{(i)})^T||_F^2 + \sum_{g=1}^{q} b_g||X^{(g)} - W^{(g)}(W^{(g)})^T||_F^2 + \sum_{t=1}^{p} c_t||(U^{(i)})^TU^{(i)} - S^TS||_F^2 \\
& + \sum_{g=1}^{q} d_g(W^{(g)})^T(W^{(g)}) - S^TS||_F^2 \quad s.t. \ U^{(i)}_{ij} \geq 0, \ W^{(g)}_{ij} \geq 0, \ S_{ij} \geq 0, \ \forall i, j,
\end{align*}$$

(1)

where $a_t, b_g, c_t, d_g$ play the role of multidimensional association relation weight control. The elements in $S^TS$ represent cosine similarity between overlapping user community, which is understood as community similarity matrix. Compared with directly calculating the difference between community membership matrix $U^{(i)}(W^{(g)})$ and result matrix $S$, calculating the difference between similarity matrix $(U^{(i)})^TU^{(i)}(W^{(g)})^T(W^{(g)})$ and similarity matrix $S^TS$ can effectively reduce the influence of noise data. Other parameters are described in Table 1.

### Optimization

To solve the matrix $S$, it is required that the difference between the matrix $S$ and the user-community membership matrix under all relations is the minimum.

Aiming at the optimization objective function given in Equation (1), the Lagrange multiplier method is used to infer $U^{(i)}$, $H^{(i)}$, $W^{(g)}$, and $S$ respectively. At the same time, owing to the element value range is different between matrix $U^{(i)}$ and $W^{(g)}$, the nonzero diagonal matrices $Q^{(i)}$ and $R^{(g)}$ are introduced to normalize each row element to the value range $[0,1]$ to guarantee each row's element sum is 1. Here $Q^{(i)} = \text{Diag}(1/\sum_{i=1}^{k} U^{(i)}, 1/\sum_{i=1}^{k} U^{(i)}, \ldots, 1/\sum_{i=1}^{k} U^{(i)})$, $R^{(g)} = \text{Diag}(1/\sum_{g=1}^{l} W^{(g)}, 1/\sum_{g=1}^{l} W^{(g)}, \ldots, 1/\sum_{g=1}^{l} W^{(g)})$.

Based on Equation (1), the Lagrange function is first given as follows:

$$\begin{align*}
\min_{U^{(i)}, W^{(g)}, S} & \sum_{t=1}^{p} (a_t||A^{(t)} - U^{(i)}H^{(i)}(U^{(i)})^T||_F^2 + c_t||(Q^{(i)}U^{(i)})^T(Q^{(i)}U^{(i)}) - S^TS||_F^2 - tr(a_t(U^{(i)})^T - tr(\beta_t(H^{(i)})^T)) \\
& + \sum_{g=1}^{q} (b_g||X^{(g)} - W^{(g)}(W^{(g)})^T||_F^2 + d_g||(R^{(g)}W^{(g)})^T(R^{(g)}W^{(g)}) - S^TS||_F^2 - tr(\gamma^{(g)}(W^{(g)})^T)) - tr(\lambda S^T) \\
& s.t. \ U^{(i)}_{ij} \geq 0, \ W^{(g)}_{ij} \geq 0, \ S_{ij} \geq 0, \ (\alpha^{(i)})_{ij}(U^{(i)})_{ij} = 0, (\beta^{(i)})_{ij}(H^{(i)})_{ij} = 0, \ (\gamma^{(g)})_{ij}(W^{(g)})_{ij} = 0, \ \lambda S_{ij} = 0.
\end{align*}$$

(2)

### Table 1: CMN_NMF symbol description

| Symbol | Description | Dimension |
|--------|-------------|-----------|
| $p$    | The number of directed network relations | -         |
| $q$    | The number of undirected network relations | -         |
| $k$    | The number of communities in integrated network | -         |
| $n$    | The number of overlapping users | -         |
| $A^{(t)}$ | User adjacency matrix based on the $t$-th directed relation | $n \times n$ |
| $U^{(i)}$ | User-community membership matrix based on the $t$-th directed relation | $n \times k$ |
| $H^{(i)}$ | Community-community relation matrix based on the $t$-th directed relation | $k \times k$ |
| $X^{(g)}$ | Semantic similarity matrix of content topics among users in social network $g$ | $n \times n$ |
| $W^{(g)}$ | User-community membership matrix based on semantic similarity of text topics in social network $g$ | $n \times k$ |
| $S$    | User-Community membership matrix based on fusion of all network relations | $n \times k$ |
Then, let \( \frac{\partial L}{\partial U} = 0, \frac{\partial L}{\partial H} = 0, \frac{\partial L}{\partial W} = 0, \frac{\partial L}{\partial S} = 0 \), and following the KKT condition \( (a^{(i)})_j(U^{(i)})(H^{(i)})_j = 0, (\beta^{(j)})_i(H^{(i)})_j = 0, (\gamma^{(q)})_i(W^{(q)})_j = 0, \lambda_i S_i = 0 \), the updating rules can be obtained, which are Equation (3), Equation (4), Equation (5), and Equation (6) respectively.

\[
U^{(i)} \leftarrow U^{(i)} \odot \frac{a_i \{ (A^{(i)} U^{(i)})(H^{(i)})^T + (A^{(i)})^T U^{(i)} H^{(i)} + 2c_i (Q^{(i)} Q^{(i)} U^{(i)} S) \} + \frac{a_i \{ (U^{(i)} H^{(i)}(U^{(i)})^T U^{(i)} H^{(i)})^T + U^{(i)}(H^{(i)})^T (U^{(i)} H^{(i)}) + 2c_i (Q^{(i)} Q^{(i)} U^{(i)}(Q^{(i)} U^{(i)})^T Q^{(i)} U^{(i)}) \}}{a_i \{ (U^{(i)} H^{(i)}(U^{(i)})^T U^{(i)} H^{(i)})^T + U^{(i)}(H^{(i)})^T (U^{(i)} H^{(i)}) + 2c_i (Q^{(i)} Q^{(i)} U^{(i)}(Q^{(i)} U^{(i)})^T Q^{(i)} U^{(i)}) \}}, \tag{3}
\]

\[
H^{(i)} \leftarrow H^{(i)} \odot \frac{(U^{(i) T} A^{(i)} U^{(i)})}{(U^{(i) T} U^{(i)} H^{(i)}(U^{(i)})^T U^{(i)})}, \tag{4}
\]

\[
W^{(q)} \leftarrow W^{(q)} \odot \frac{b_q X^{(q) W^{(q)}} + d_q (R^{(q)})^T R^{(q)} W^{(q)} S^T}{b_q W^{(q) T} W^{(q)}} + (R^{(q)})^T R^{(q)} W^{(q)} (R^{(q)} W^{(q)})^T R^{(q)} W^{(q)}}, \tag{5}
\]

\[
S \leftarrow S \odot \frac{\sum_{i=1}^{p} c_i S(Q^{(i) U^{(i)}} T Q^{(i) U^{(i)}} + \sum_{q=1}^{q} c_i S(Q^{(i) U^{(i)}} T Q^{(i) U^{(i)}})}{\sum_{i=1}^{p} c_i S(S) + \sum_{q=1}^{q} d_q S(S)}, \tag{6}
\]

Based on the iterative updating rules, CMN_NMF algorithm can be described as Algorithm 1.

**Algorithm 1. CMN_NMF algorithm**

**Input:** \( p \) asymmetric user adjacency matrices, \( q \) symmetric user adjacency matrices, the number of stub communities \( k \)

**Output:** stub community membership matrix \( S \)

1. Initialize the matrix \( U^{(0)}, H^{(0)}, W^{(0)}, S \)
2. **while** Equation (1) not converge **do**
3. \( \text{for } t = 1 \text{ to } p \text{ do} \)
4. Update \( U^{(t)} \) according Equation (3)
5. Update \( H^{(t)} \) according Equation (4)
6. **end for**
7. **for** \( t = 1 \text{ to } p \text{ do} \)
8. Update \( W^{(q)} \) according Equation (5)
9. Update \( S \) according Equation (6)
10. **end for**
11. **end while**

### 4.1.4 Convergence analysis

The iterative formula of parameters given in this study can guarantee the final convergence of the algorithm. The following is the proof of the convergence of \( U^{(t)} \); this method is also applicable to other parameters.

Let 
\[
L(U^{(t)}) = a_0 \| A^{(i)} - U^{(t)} H^{(t)}(U^{(t)})^T \|_F^2 + c_i \| (U^{(t)})^T S^{(i)} T - S^{(i)} T S \|_F^2. \tag{7}
\]

Now \( L(U^{(t)}) \) is convex function with respect to \( U^{(t)} \), and we need to prove that \( L(U^{(t)}) \) is nonincreasing function. That is:

\[
L((U^{(t)})^{m+1}) \leq L((U^{(t)})^m), \tag{8}
\]

where \( U^{(t)} \) represents that the \( U^{(t)} \) is obtained after the \( x \)-th iteration with Equation (3).
Definition: \( Z(U^{(i)}, U^{(j)}) \) is the auxiliary function of \( L(U^{(i)}) \), which meets the conditions as follows:

\[
Z(U^{(i)}, U^{(j)}) \geq L(U^{(i)}),
Z(U^{(i)}, U^{(i)}) = L(U^{(i)}).
\]

Proof:

\[
L((U^{(i)})^{m+1}) = Z((U^{(i)})^{m+1}, (U^{(i)})^m) \leq Z((U^{(i)})^m, (U^{(i)})^m) = L((U^{(i)})^m).
\]

First, the auxiliary function \( Z \) is given:

\[
Z(U^{(i)}, U^{(j)}) = L(U^{(j)}) - 2 \left( a_i (A^{(i)} U^{(j)} H^{(i)} T + (A^{(i)})^T U^{(j)} H^{(i)}) + 2c_i (Q^{(i)} Q^{(j)} U^{(j)} S^T S) U^{(j)} (\log U^{(i)} - \log U^{(j)}) + 2c_i (Q^{(j)} Q^{(i)} U^{(j)} U^{(i)} S^T S) U^{(i)} (\log U^{(i)} - \log U^{(j)}) + 2c_i (Q^{(i)} Q^{(j)} U^{(j)} (Q^{(i)} U^{(i)} S^T S) U^{(i)}) \left( \frac{(U^{(i)})^2 + (U^{(j)})^2}{2U^{(j)}} - U^{(j)} \right) \right).
\]

The Taylor expansion of the \( L(U^{(i)}) \) function is compared with the auxiliary function \( Z \).

When \( U^{(j)} = U^{(i)}, Z(U^{(i)}, U^{(i)}) = L(U^{(i)}); \)

When \( U^{(j)} \neq U^{(i)}, Z(U^{(i)}, U^{(j)}) \geq L(U^{(i)}), \) we only need to prove that:

1. \(- (U^{(i)} - U^{(j)}) \leq -U^{(j)} (\log U^{(i)} - \log U^{(j)}) \)
2. \( U^{(i)} \leq \frac{(U^{(i)})^2 + (U^{(j)})^2}{2U^{(j)}} \)

The conclusions of these are easy to be proved, and \( Z(U^{(i)}, U^{(j)}) \geq L(U^{(i)}) \) is established, so that:

\[
\frac{\partial Z(U^{(i)}, U^{(j)})}{\partial x} = 0.
\]

After \( U^{(i)} \) is obtained in terms of \( U^{(j)} \), we can update \( U^{(i)} \) according to rule, that is, Equation (3). Other iterative updating formulas are the similar as above.

### 4.2 Reconstructing single social network

After obtaining all overlapping user stub communities divided in the heterogenous integrated network, inspired by seed based community detection methods, this paper makes use of stub communities and expand them in different social networks. Each stub community includes users in single social network who meet certain condition described below. However, the number of users in each social network is large which may lead to high computing complexity, so we only utilize users’ “follow” relation which represents information spread direction to reconstruct each social network. By this way, the computing cost is reduced and the scalability of the algorithm is improved.

The new reconstructed network is a weighted undirected network. The edge in network represents the two users have overlapping followees. We define \( F_i \) and \( F_j \) as user \( i \) and \( j \)’s followee sets, respectively. The weight of edge (similarity between users) is calculated by Jaccard Similarity as follows:

\[
\text{Sim}_{ij} = \frac{\text{intersection of user } i, j’s \text{ followees}}{\text{union of user } i, j’s \text{ followees}} = \frac{|F_i \cap F_j|}{|F_i \cup F_j|}. \tag{10}
\]
Algorithm 2. Reconstruction method of single social network

**Input:** $G^i_{\text{old}} = (U^i, E^i_{\text{follow}})$, $U^i$: represents all users of $i$-th network, $E^i_{\text{follow}}$: represent all edges of “follow” relation in $i$-th network

**Output:** $G^i_{\text{new}} = (U^i, E^i_{\text{sim}})$, $U^i$: represents all users of $i$-th network, $E^i_{\text{sim}}$: represent all edges of user similarity in $i$-th network

1. $E^i_{\text{sim}} \leftarrow \emptyset$
2. for each $i \in U^i$ do
3.   for each $j \in U^i$ do
4.     $\text{Sim}_{ij} = \frac{|F_i \cap F_j|}{|F_i \cup F_j|}$
5.     if $\text{Sim}_{ij} > \text{threshold value}$ then
6.       $E^i_{\text{sim}}$ add a edge between $i$ and $j$
7.     endif
8.   endfor
9. endfor

### 4.3 Community detection algorithm based on overlapping user stub community

In this study, overlapping user stub community set $N^0 = \{N^0_1, N^0_2, \ldots, N^0_k\}$ is obtained by community detection on integrated network, and $G^i_{\text{new}} = (U^i, E^i_{\text{sim}})$ is reconstructed based on original $i$-th network according to users’ “follow” relations. In order to determine users who can be added in community, the community connection strength $\text{NS}_{iN^0_i}$ is defined to measure the closeness between user $i$ and stub community $N^0_i$. And individual connection strength $\text{NS}_{ij}$ is defined to represent the maximum value of all paths’ values between user $i$ and user $j$. The path value is obtained by multiplying every edge’s weight in this path. Then, we can get $\text{NS}_{iN^0_i}$ calculation formula as follows:

$$\text{NS}_{iN^0_i} = \frac{\sum_{j \in N^0_i} \text{NS}_{ij}}{\text{size}_{N^0_i}}.$$  \hspace{1cm} (11)

The threshold value $\text{Rec}_i$ is defined as condition to include new user into stub community $N^0_i$, that is, if $\text{NS}_{iN^0_i} > \text{Rec}_i$, the user is put in community $C_{i,new}$ which is expanded by stub community $N^0_i$. The community detection algorithm based on overlapping user stub community is shown as Algorithm 3. We can get expanded communities of every stub community on $G_{\text{new}}$. In the same way, the stub communities could be expanded in other social networks. Finally, the cross-social network communities are obtained by integrating all expansions.

Algorithm 3. Community detection algorithm based on overlapping user stub community

**Input:** User set $U^i$ in reconstructed $i$-th network $G^i_{\text{new}}$, Connection strength set $\text{NS}_{iN^0_i}$

**Output:** Community set $C_{\text{new}}$ in reconstructed $i$-th network $G^i_{\text{new}}$

1. for each $N^0_i \in N^0$ do
2.   for each $u_i \in (U - N^0_i)$ do
3.     if $\text{NS}_{iN^0_i} > \text{Rec}_i$ then
4.       $C_{i,new}.\text{add}(u_i)$
5.     endif
6.   endfor
7. endfor
8. return $C_{\text{new}} = \{C_{1,new}, C_{2,new}, \ldots, C_{k,new}\}$
**Table 2** Overlapping user data statistics

| Overlapping User Types of Relation | Number |
|-----------------------------------|--------|
| The “follow” relation             | 982    |
| The “comment” relation            | 2065   |
| The “like” relation               | 6243   |

**Table 3** Non-overlapping user data statistics

| Social Network | The Number of Users | The Number of Relations |
|----------------|--------------------|-------------------------|
| Zhihu          | 24,294             | 2,394,472               |
| Weibo          | 180,736            | 5,377,427               |

**Table 4** Average number of communities and modularity based on FU algorithm

| The Type of Subnetwork | The Number of Communities | Modularity |
|------------------------|---------------------------|------------|
| “follow” relation      | 15.6                      | 0.38       |
| “comment” relation     | 17.4                      | 0.49       |
| “like” relation        | 15.1                      | 0.27       |

Abbreviation: FU, fast-unfolding.

5 | EXPERIMENTS AND ANALYSIS

5.1 | Data description

In this study, 1575 overlapping user accounts come from Zhihu-Weibo dataset. Here, “Zhihu-Weibo” is a hybrid data including overlapping users’ data from Zhihu and Weibo network. Zhihu is a question-answer based social network, and Weibo is a microblog-based social network. A set of users, together with their relations, profiles, and textual contents, was firstly crawled from Zhihu, then the linked Weibo accounts were obtained from Zhihu users’ profiles, finally we further crawled these users’ Weibo data. The data statistics are shown in Table 2.

Based on the above overlapping user set, the additional nonoverlapping users can be obtained by users’ followees respectively in Zhihu and Weibo networks, the statistics on “follow” relations among nonoverlapping user dataset are shown in Table 3.

5.2 | Experiments design

5.2.1 | The number of communities

The number \( k \) of communities is a key parameter as an input of the algorithm CMN_NMF, the reasonable setting of \( k \) value plays an important role in the quality of community detection. This study uses the fast-unfolding (FU) algorithm\(^6\) to determine the \( K \) value range. FU algorithm is based on topology structure by optimizing modularity to obtain the most appropriate number of community detections.

FU algorithm can hardly handle fused multidimensional attributed network. We only use this algorithm to determine the number of community detections in each type of subnetwork. Here, we run FU algorithm 10 times for each type of subnetwork, that is, \( G_1^c, G_2^c, G_3^c \) to get the optimal parameter \( k \) by computing the average modularity. The details are shown in Table 4.

It can be seen from Table 4 that about 16 communities are best divided based on FU algorithm for each subnetwork in integrated network. In order to ensure the experimental result reliability, the value range of \( k \) is from 14 to 20.

5.2.2 | Evaluation metrics

There are two common evaluation metrics of community detection in complex network. The first is applied when actual members of community are already known in the network. The evaluation metric is to compare the difference between
detected community and actual community. In social networks, it is usually difficult to identify the members of community in the network. The second is to measure the community detection results by calculating the network modularity. The modularity is widely used to evaluate early 1D attributed community detection works, and we have already used modularity to determine the number of community detections. Research objective in this paper is to find similar users across social networks. Based on the above considerations, the following two evaluation metrics are given in this paper.

**Similarity of user textual content.** Similar users have similar content interest. These similar users in different social networks are derived based on overlapping user stub communities, and validity of algorithm can be verified by calculating the similarity of users’ post content from different social network in the same community. Here we take Weibo and Zhihu data sets as examples. $C_1$ represents cross-social network community derived from Weibo and Zhihu networks based on overlapping stub community $N_1^0$. $C_1^c$ represents user set from Weibo social network in community $C_1$ (excluding stub community users), and $C_1^z$ represents user set from Zhihu social network in community $C_1$ (excluding stub community users). If the Weibo user set and Zhihu user set in same community have a high similarity, then the textual content of the two user sets should have a high consistency. This paper obtains textual content from Weibo and Zhihu networks respectively, calculates the word vectors of users’ post content in two user sets based on TF-IDF, and measures consistency based on cosine similarity.

**Implicit overlapping user discovery.** The clustering consensus feature\(^{16,17}\) of overlapping users under different networks is used to detect community. Referring to this concept of feature, we propose a new evaluation metric, implicit overlapping user discovery ratio, to reflect the consensus of overlapping user cluster across different social networks in this paper. For example, $u_i$ and $u_j$ are two overlapping users in Weibo and Zhihu, if $u_i$ and $u_j$ are in one community in Weibo, then they also tend to be in one community in Zhihu. This paper divides overlapping users into training users and test users. The training users are used to compose stub community structure, and test users are distributed in each social network for further verifying discovery effectiveness of implicit overlapping users.

The formal expression of implicit overlapping user discovery ratio is as follows:

\[
P = \frac{|C_{\text{overlapping}}|}{|U_{\text{test overlapping}}|},
\]

where $|C_{\text{overlapping}}|$ represents the number of implicit overlapping users who are part of test overlapping users and $|U_{\text{test overlapping}}|$ is the number of test overlapping users.

### 5.2.3 Comparison methods

In order to verify the effectiveness of the cross-social network community detection method based on overlapping users proposed in this study, our method was compared with other methods. The methods involved in the experiment are introduced as follows:

**K-means:** a fusion network was obtained by fusing the multidimensional relations of the overlapping user integrated network, and the k-means clustering algorithm was used to find stub community.

**ConcatNMF:** a fusion network was obtained by fusing the multidimensional relations in the overlapping user integrated network, and the NMF clustering algorithm was used to find stub community.

**ColNMF:**\(^{21}\) a shared result matrix $S$ is introduced, which is stub community membership matrix. NMF clustering algorithm was used to simultaneously find stub communities in each relation network by using one shared matrix $S$. The formula is as follows:

\[
\min \sum_{v=1}^{n} \lambda_v \|A^{(v)} - U^{(v)}S^T\|_F^2.
\]

**MultiNMF:**\(^{13}\) a shared result matrix $S$ is arranged, which is stub community membership matrix. NMF clustering algorithm was used to find stub communities in each relation network respectively, and shared result matrix $S$ is used to associate each stub community matrix in every relation network. The formula is as follows:

\[
\min \sum_{v=1}^{n} \|A^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n} \lambda_v \|V^{(v)} - S\|_F^2.
\]
5.3 Results analysis

When $k$ value is 17, cross-social network community detection results based on overlapping users are obtained. Users are divided into 17 communities, and the communities were sorted by the number of overlapping users from high to low. The similarity between Weibo users’ textual content and Zhihu users’ textual content in each community was calculated. The results are shown in Figure 2.

When $k$ value is calculated as 14 to 20, the average textual content similarity of Weibo and Zhihu users in the communities is calculated in each $k$ value. The results are shown in Figure 3.

It can be seen from the figure that when the number of community detections is greater than 16, the average similarity of textual content of Weibo and Zhihu users in all communities gradually tends to be stable.

The average similarity of users’ textual content is shown in Figure 4, which is calculated and compared with methods described above.

This figure shows the effectiveness of the stub community detection algorithm proposed in the cross-social network community detection. Although MultiNMF has similar solution form based on NMF clustering algorithm compared with our method, it shows worse performance due to ignorance of textual content specific to topic. The basic ideas of Co1NMF, ConcatNMF, and k-means algorithms are to directly fuse different types of user relations before stub community detection. Data noise and dimensional difference, and other problems have a significant impact on the results.

**FIGURE 2** Similarity of textual content between Weibo and Zhihu users in each community

**FIGURE 3** Average similarity of users textual content with different $k$ values

**FIGURE 4** Comparison of average similarity of users textual content with different methods
In this paper, one-third of overlapping users were randomly selected as test users, and stub communities are composed of remaining two-third of overlapping users in Weibo and Zhihu. When $k$ is set to 17, the discovery ratio of implicit overlapping users in each expanded community is shown in Figure 5.

When $k$ value is set as 14 to 20, the average implicit overlapping user discovery ratio in community is shown in Figure 6. The decrease of average ratio reflects that the discovery efficiency of implicit overlapping users will be affected by the increase of the number of community detections.

The average discovery ratio implicit overlapping user is calculated, and compared with other algorithms. The results are shown in Figure 7. The results demonstrate the effectiveness of integrated network stub community detection algorithm in community detection, and our method is more reliable than other algorithms to find overlapping users.

5.4 Conclusion

In order to detect user community across social networks, this study starts from overlapping users across social networks and finds similar users in different social networks through overlapping user stub communities based on the basic
idea that most overlapping users have similar friends in different social networks. In the process of finding overlapping user stub communities, considering relation data sparsity in overlapping users, this paper firstly takes advantage of non-negative matrix decomposition algorithm to get user stub communities by integrating users’ different relations. These relations, including user link relations and content relations, effectively improve the consistency to guarantee the effectiveness of the overlapping users stub community detection. Then, this paper presents a simple and effective method to reconstruct single social network. Finally, a social network community detection method based on overlapping user stub communities is proposed. This method determines the community by calculating the connection strength between users and overlapping user stub communities in each social network. The community size can be flexibly adjusted through threshold modification, which has good adaptability. Experiments show that our method is effective in community detection.

Further works will focus on two aspects. First, data from heterogeneous social networks such as images and locations can be involved to carry out cross-social network community detection in heterogeneous social networks. Second, the time factor can be used to discuss the dynamic evolution on cross-social network communities.

ACKNOWLEDGMENTS
This work is supported by National Natural Science Foundation of China under Grants No. 61772133 and No. 61972087, National Social Science Foundation of China under Grants No. 19@ZH014, Jiangsu Provincial Key Project under Grants No. BE2018706, Natural Science Foundation of Jiangsu province under Grants No. SBK2019022870, Jiangsu Provincial Key Laboratory of Computer Networking Technology, Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No. BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No. 93K-9.

REFERENCES
1. Xiang L, Sang J, Xu C. Cross-OSN user modeling in big data. Big Data Res. 2016;2(05):32-42.
2. Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci. 2002;99(12):7821-7826.
3. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004;69(2):026113.
4. Newman ME. Fast algorithm for detecting community structure in networks. Phys Rev E. 2004;69(6):066133.
5. Clauset A. Finding local community structure in networks. Phys Rev E. 2005;72(2):026132.
6. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):P10008.
7. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435(7043):814.
8. Derényi I, Palla G, Vicsek T. Clique percolation in random networks. Phys Rev Lett. 2005;94(16):160202.
9. Lanchinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. New J Phys. 2009;11(3):033015.
10. Wang F, Li T, Wang X, Zhu S, Ding C. Community discovery using nonnegative matrix factorization. Data Min Knowl Disc. 2011;2:493-521.
11. Zhao Z, Feng S, Wang Q, Huang JZ, Williams GJ, Fan J. Topic oriented community detection through social objects and link analysis in social networks. Knowl-Based Syst. 2012;26:164-173.
12. Xu S, Cao J, Zhu X, Dong Y, Liu B. Community discovery based on social relations and temporal-spatial topics in LBSNs. Paper presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2018:206-217; Springer, Cham.
13. Liu J, Wang C, Gao J, Han J. Multi-view clustering via joint nonnegative matrix factorization. Paper presented at: Proceedings of the 2013 SIAM International Conference on Data Mining; 2013:252–260; SIAM.
14. He X, Kan MY, Xie P, Chen X. Comment-based multi-view clustering of web 2.0 items. Paper presented at: Proceedings of the 23rd International Conference on World Wide Web (WWW ’14); 2014:771-782; ACM, New York, NY.
15. Pei Y, Chakraborty N, Sycara K. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. Paper presented at: Proceedings of the 24th International Joint Conference on Artificial Intelligence; 2015.
16. Nguyen HT, Dinh TN, Vu T. Community detection in multiplex social networks. Paper presented at: Proceedings of the 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); 2015:654-659; IEEE.
17. Philip SY, Zhang J. MCD: mutual clustering across multiple social networks. Paper presented at: Proceedings of the 2015 IEEE International Congress on Big Data; 2015:762-771; IEEE.
18. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. Adv Neural Inf Process Syst. 2001;13:556-562.
19. Zhu L, Galstyan A, Cheng J, Lerman K. Tripartite graph clustering for dynamic sentiment analysis on social media. Paper presented at: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data; 2014:1531-1542; ACM.
20. Pan L, Jin J, Wang CJ, Xie JY. Detecting link communities based on local information in social networks. Dianzi Xuebao (Acta Electronica Sinica). 2012;40(11):2255-2263.
21. Singh AP, Gordon GJ. Relational learning via collective matrix factorization. Paper presented at: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2008:650-658; ACM.

**How to cite this article:** Zhu Z, Zhou T, Jia C, Liu W, Liu B, Cao J. Community detection across multiple social networks based on overlapping users. *Trans Emerging Tel Tech*. 2022;33:e3928. [https://doi.org/10.1002/ett.3928](https://doi.org/10.1002/ett.3928)