PAPER

Hierarchical Community Detection in Social Networks Based on Micro-Community and Minimum Spanning Tree

Zhixiao WANG[†], **Mengnan HOU**[†], **Guan YUAN**[†a)], **Jing HE**[†], **Jingjing CUI**[††], ***Nonmembers***,
and **Mingjun ZHU**[†], ***Member***

SUMMARY Social networks often demonstrate hierarchical community structure with communities embedded in other ones. Most existing hierarchical community detection methods need one or more tunable parameters to control the resolution levels, and the obtained dendrograms, a tree describing the hierarchical community structure, are extremely complex to understand and analyze. In the paper, we propose a parameter-free hierarchical community detection method based on micro-community and minimum spanning tree. The proposed method first identifies micro-communities based on link strength between adjacent vertices, and then, it constructs minimum spanning tree by successively linking these micro-communities one by one. The hierarchical community structure of social networks can be intuitively revealed from the merging order of these micro-communities. Experimental results on synthetic and real-world networks show that our proposed method exhibits good accuracy and efficiency performance and outperforms other state-of-the-art methods. In addition, our proposed method does not require any pre-defined parameters, and the output dendrogram is simple and meaningful for understanding and analyzing the hierarchical community structure of social networks.

key words: *social network analysis, hierarchical community detection, micro-community, minimum spanning tree*

1. Introduction

Many social networks exhibit a natural community structure, i.e., groups of vertices that have denser connections within each group and fewer connections between them [1]. Community detection is a hot issue in social network analysis, which is beneficial for us to analyze the topology structure of social networks, understand the evolution of social networks, and even forecast their behaviors [2].

The community structure of social networks often demonstrates hierarchy with small communities embedded in big ones [3]. Many hierarchical community detection methods were proposed [4]–[7]. These methods can be classified into two categories: agglomerative [8]–[10] and divisive [11], [12]. The former is a bottom up approach that successively joins the initial vertices while the latter divides the network into sub-groups. Because of its excellent performance, agglomerative method has attracted many attentions. Typical agglomerative algorithms include: modularity optimization [13]–[16], density-based methods [17]–[21], game theory-based methods [22], [23], random walk-based methods [24]–[26], and NMF (Nonnegative Matrix Factorization)-based methods [27], [28]. These methods have achieved good performance in community detection, but there are still some defects among them. For example, the modularity optimization methods have a problem of minimum resolution, the game theory-based methods need some pre-defined parameters to adjust the context of the coalition.

Although many hierarchical community detection methods have been proposed based on different technologies, there still remain some unsolved problems: **(1)** The obtained dendrograms of social network communities are extremely complex with very large depth. From these complicated dendrograms, we cannot efficiently analyze the hierarchical relationship among communities. **(2)** Most existing hierarchical community detection methods need one or more pre-defined tunable parameters to control the resolution levels. In fact, it is difficult to set appropriate values for different social networks in advance. Thus, the partition performance cannot be effectively guaranteed.

The fundamental reason of the first problem is that most existing methods successively merge the single vertices one by one to get hierarchical structure. The bigger the social network scale is, the more complex the obtained dendrogram will be. The bottom level of dendrogram refers to single vertices. These single vertices are too tiny to be meaningful for understanding the hierarchical community structure of social networks. If we can construct the dendrogram from local communities, rather than single vertices, the obtained dendrogram will become simple and intuitive.

In order to solve the second problem, some parameter-free methods need to be adopted. The Minimum Spanning Tree (MST) is a kind of parameter-free agglomerative method used in hierarchical clustering. Although some methods have utilize MST to identify community structure of social networks [29]–[31], these method mainly focused on how to remove some edges from the constructed MST, based on pre-defined rules, to obtain community structure. To the best of our knowledge, none of them devoted to mining the hierarchical relationship among communities. If we construct MST from local communities, these local communities will be successively linked (i.e. merged) one by one and the merging order will reveal how these local communities are nested in different levels. Thus, we can identify

Manuscript received June 9, 2018.
Manuscript revised March 21, 2019.
Manuscript publicized June 5, 2019.

[†] The authors are with School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China.

[††] The author is with Baidu Online Network Technology (Beijing) Co., Ltd, Beijing, 100085 P.R. China.

a) E-mail: yuanguan@cumt.edu.cn

DOI: 10.1587/transinf.2018EDP7205

the hierarchical relationship among communities of social networks.

Motived by above inspirations, we proposed a novel hierarchical community detection method based on micro-community and minimum spanning tree. Firstly, the proposed method identifies dense pairs from social networks. A dense pair is a pair of vertices with the largest link strength from each other. Secondly, all dense pairs sharing common vertices are merged into bigger local communities, each local community is called a micro-community. Thirdly, the proposed method constructs MST by successively linking these micro-communities one by one. The merging order of these micro-communities will intuitively reveal the hierarchical community structure of social networks. Experimental results show that our proposed method exhibits good performance on synthetic and real-world networks and outperforms other state-of-the-art methods. The main contributions of this paper are summarized as follows:

(1) The proposed method constructs the dendrogram from identified micro-communities, rather than single vertices. Thus, the obtained dendrogram can intuitively reveal how the micro-communities are nested in different levels, which is meaningful for understanding and analyzing the hierarchical community structure of social networks.

(2) The proposed method directly detects the hierarchical community structure of social networks in the process of minimum spanning tree construction, requiring no pre-defined tunable parameters or edge removing rules. Thus, the partition efficiency can be effectively guaranteed.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 describes the proposed method in detail. Section 4 discusses the experimental results and Sect. 5 provides the conclusion of this paper.

2. Related Works

Agglomerative hierarchical community detection methods can be divided into several categories, including modularity optimization [13]–[16], density-based methods [17]–[21], game-theory-based methods [22], [23], random-walk-based methods [24]–[26], and NMF-based methods [27], [28].

Modularity optimization. FN [8] and CNM [9] are two representative modularity optimization methods. FN [8] starts with a state in which each vertex is the sole member of each community, and then repeatedly merge community pairs that result in the greatest increasing in Q. The order of the joining of a pair of communities can be represented as a dendrogram, a tree describing the hierarchical community structure. CNM [9] is another modularity-based method that obtains the hierarchical community structure by exploiting some shortcuts in the optimization problem and using more sophisticated data structures. Its running time on a network with n vertices is $O(nlog^2n)$. Toujani and Akaichi [13] propose a hybrid method to uncover the hierarchical community structure of complex networks. However, the method assumes the existence of an initial partition, which has significance influence to the final results. Afterwards, they improve the method by presenting an objective function [14] that incorporates the value of structural and semantic similarity based on modularity. Barber et al. [15] describe a hierarchical clustering algorithm that merges all pairs of communities connected by locally optimal edges that will increase the modularity. The worst-case time complexity is $O(n^3)$, which isn't suitable for large scale networks. Lin et al. [16] propose an Integer Programming approach IP, a bottom-up method for detecting hierarchical community structure. Whereas, this approach needs a pre-defined hierarchical level.

Density-based methods. Huang et al. [17] propose a hierarchical network clustering method DenShrink to reveal the embedded hierarchical community structure by combining the advantage of density-based clustering and modularity optimization. DenShrink identifies micro-communities based on structural similarity between adjacent vertices, and then merge the densely connected local micro-communities iteratively to reveal the hierarchical community structure.

Game-theory-based methods. Zhou et al. [22] propose a game-theory-based method to identify hierarchical communities, which models individual vertices as rational players. The players cooperate with other players to form coalitions, and coalitions with fewer players can merge into a larger coalition as long as the merge operation can improve the utility value of the merged coalitions. The process of merging coalitions illustrates the forming of the hierarchical communities. However, how to select a suitable value for β, a parameter used to adjust the context of the coalition, is a challenge task.

Random-walk-based methods. Zhang et al. [24] propose a hierarchical community detection method based on random walk, which utilizes the error function of the partial transition matrix convergence to determine the number of random walk steps. The method clusters the communities around core vertices based on the approximation convergence transition matrix, and then uses a closeness index to recursively merge two communities. However, this method needs a pre-defined closeness threshold φ to decide the mergence of two communities, and a threshold τ to select the initial core vertex set.

NMF-based methods. Du et al. [27] propose a NMF-based method, they provide a divide-and-conquer strategy to discover hierarchical community structure based on the rank-2 symmetric nonnegative matrix factorization. However, when splitting a community using rank-2 SymNMF, an appropriate scalar parameter for the tradeoff between the approximation error and the difference between matrix H and W is needed.

The dendrograms obtained by most above methods are extremely complex with very large depth. Efficiently analyzing the hierarchical relationship of communities from these complicated dendrograms is difficult. Furthermore, most existing hierarchical community detection methods need one or more pre-defined tunable parameters. Setting appropriate values for different social networks in advance is also difficult.

3. Method

The paper proposes a parameter-free hierarchical community detection method based on micro-community and minimum spanning tree. The method includes three steps. The first step is dense pair identification. A dense pair is a pair of adjacent vertices with the largest link strength from each other. The second step is micro-community generation. All identified dense pairs sharing common vertices are merged into bigger local communities, each local community is called a micro-community. The third step is MST construction. Prim algorithm is used to generate minimum spanning tree, and the generated micro-communities are regarded as initial vertices. The hierarchical community structure of social networks can be intuitively revealed by successively linking these micro-communities one by one. The root (highest level) represents a single community, corresponding to the entire network and the leaves (lowest level) corresponds to the micro-communities.

Definition 1 (Link Strength). For a given network $G = (V, E)$, V denotes the vertices set, E represents the edges set. For $\forall u \in V$ and $v \in \Gamma(u) - \{u\}$ where $\Gamma(u)$ represents the neighbor set of vertex u including u itself, the Link Strength between vertex u and v is defined as follows:

$$LS(u, v) = \left|\frac{\Gamma(u) \cap \Gamma(v)}{\Gamma(u) \cup \Gamma(v)}\right| + \frac{\left|\frac{D(u)}{\sum_{l \in \Gamma(v)} D(l)} - \frac{D(v)}{\sum_{l \in \Gamma(u)} D(l)}\right|}{\frac{1}{2} \times \left(\frac{D(u)}{\sum_{l \in \Gamma(v)} D(l)} + \frac{D(v)}{\sum_{l \in \Gamma(u)} D(l)}\right)} \quad (1)$$

where, $LS(u, v)$ represents the Link Strength between vertex u and v. $D(u)$ denotes the degree of vertex u. The traditional network similarity (such as Jaccard similarity) calculates the similarity of two vertices based on the link relationship of their neighbourhood (i.e. one hop). However, the proposed Eq. (1) not only takes the information of one hop neighbourhood into account, but also considers the information of neighbour's neighbour (i.e. two hops), resulting in a more holistic and accurate result. That is to say, the front part of Eq. (1) calculates the Link Strength from the perspective of one hop, and the latter evaluates it from the perspective of two hops.

Definition 2 (Dense Pair). Given a network $G = (V, E)$. For $\forall u \in V$ and $v \in \Gamma(u) - \{u\}$ where $\Gamma(u)$ refers to the neighbor set of vertex u including u itself, (u, v) is a dense pair if it satisfies: $LS(u, v) = \max\{LS(x, y) | x = u, y = \Gamma(u) - \{u\}\}$ where $LS(u, v)$ denotes the Link Strength of vertex u and v.

Definition 3 (Micro-community). Given a network $G = (V, E)$. $microCom = (V', E')$ is a connected sub-graph of G, which is a micro-community iff: **(1)** for $\forall u \in V'$, $\exists v \in V'$ satisfies that (u, v) is a dense pair; **(2)** $\nexists u \in V$ that meets (u, v) is a dense pair while $u \in V' \wedge v \notin V'$.

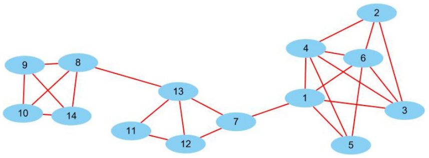


Fig. 1 A small network with three communities.

Definition 4 (Micro-community Closeness). Given a network $G = (V, E)$. C_i, C_j are two micro-communities of G. Micro-community closeness between two micro-communities C_i, C_j is defined as follows:

$$MC(C_i, C_j) = \frac{e(C_i, C_j)}{e(C_i, C_i) + e(C_j, C_j) + e(C_i, C_j)} + \left\{\frac{1}{2} \times \left(\frac{e(C_i, C_j)}{e(C_j)} + \frac{e(C_i, C_j)}{e(C_i)}\right)\right\} \quad (2)$$

where, $MC(C_i, C_j)$ represents the micro-community closeness between C_i and C_j, $e(C_i, C_j)$ refers to the number of edges connecting C_i and C_j, $e(C_i, C_i)$ denotes the number of internal edges connecting between vertices in C_i, and $e(C_i)$ represents the number of edges connecting C_i and the other micro-communities. The former part of Eq. (2) evaluates the micro-community closeness from the perspective of the internal links, and the latter indicates it from the perspective of the external links. Thus, we can evaluate the micro-community closeness holistically and accurately.

Figure 1 shows a small network with three micro-communities, we use this small example to show the effectiveness of Eq. (1) and Eq. (2).

In Fig. 1, the vertex sets $\{8, 9, 10, 14\}$, $\{7, 11, 12, 13\}$ and $\{1, 2, 3, 4, 5, 6\}$ correspond to three different micro-communities. The Link Strength between vertices 13 and 8 (they come from different micro-communities) should be smaller than that of vertices 13 and 7 (they are in the same micro-community). After calculation with Eq. (1), we get LS(13, 8) = 0.2900, LS(13, 7) = 0.6429, LS(13, 11) = 1.1484 and LS(13, 12) = 1.0800. Obviously, this result is in accordance with the intuitive judgment, indicating the effectiveness of Eq. (1). We use m1, m2, m3 to denote the micro-community $\{1, 2, 3, 4, 5, 6\}$, $\{7, 11, 12, 13\}$ and $\{8, 9, 10, 14\}$, respectively. After calculation with Eq. (2), we get MC(m1, m2) = 1.9950, MC(m1, m3) = 1 and MC(m2, m3) = 2.0054. This result also matches the reality, i.e. the micro-community closeness between m1 and m3 should be the smallest one.

```
Algorithm 1. Micro-community generation
Input: $G = (V, E)$,  $|V| = n$,  $|E| = m$
Output: the number of micro-communities $K$,
        generated micro-community $C_1, C_2, \ldots C_k$.
1) For each $e(i, j) \in E$
2)     calculate $LS(i, j)$ with Eq. (1);
3) End for
4) For $i = 1 : n$
5)     For $j = 1 : |\Gamma(i) - \{i\}|$;
```

Table 1 The characteristic and time complexity of different methods

Methods	Characteristic	Time complexity	References
Our method	Mining spanning tree	$O(m)$	–
Zhang et al	Random walk	$O(n^2 + nklogk)$	Zhang [24]
INFOMAP	Random walk	$O(m)$	Fiscarelli [10]
Gillis et al	NMF-based	$O(mnk)$	Gillis [28]
COFOGA	Game theory	$O(nlogn)$	Zhou [22]
Barber et al	Modularity optimization	$O(n^3)$	Barber [15]
GN	Betweenness centrality	$O(m^2n)$	Girvan and Newman [32]
FN	Modularity optimization	$O(n(m + n))$	Newman [8]
CNM	Modularity optimization	$o(nlog^2n)$	Clauset [9]
ANA	Mining spanning tree	$O(nlogn)$	Asmi [31]

```
6)        If $LS\ (i, j) = \max\{LS\,(x, y)|x = i, y = \Gamma(i) - \{i\}\}$
7)            $i, j$ is a dense pair, denoted as $(i, j)$
8)        End if
9)    End for
10) End for
11) $G' = (V', E')$, $V' \leftarrow V$, $E' \leftarrow DP$ //DP is the set of dense pairs.
12) Traversal $G'$ to get $K$ connected components with DFS algorithm. // DFS is the Depth-First-Search algorithm.
13) $K$ micro-communities ← $K$ connected components
```

```
Algorithm 2. Hierarchical community detection
Input: Identified micro-community set $C$,
       Micro-community number $K$.
Output: Dendrogram, including $level_1$, $level_2$, ..., $level_K$.
1) For $i$ = 1:$K$
2)    For $j = i + 1$:$K$
3)        calculate $MC(C_i, C_j)$ with Eq. (2)
4)    End for
5) End for
6) $level_1 \leftarrow K$ micro-communities
7) For $i$=2:$K$
8)    $(C_p, C_q) \leftarrow$ agrmin$(1/MC)$. // $(C_p, C_q)$ is micro-community pair with largest MC value.
9)    $level_i \leftarrow$ merge $C_p$ and $C_q$ of $level_{i-1}$
10)     MC$(C_p, C_q) = 0$
11) End for
```

Algorithm 1 identifies micro-communities of social networks, and Algorithm 2 detects hierarchical community structure based on micro-communities identified by Algorithm 1. Now, we analyze the complexity of the two algorithms.

Complexity Analysis

In Algorithm 1, steps 1−3 calculate the Link Strength between adjacent vertices, the time complexity is $O(m)$ where m represents the number of edges. Steps 4−10 identify dense pairs among vertices, the time complexity is also $O(m)$. Steps 11−13 generate micro-communities from identified dense pairs. Specifically, step 11 constructs the adjacency list of the graph $G' = (V', E')$ of dense pairs. The time complexity is $O(n)$ where n denotes the number of vertices. Steps 12−13 traverse the adjacency list of graph with the DFS (Depth First Search) algorithm to get K connected components, i.e. K micro-communities. The time complexity is $O(2n)$. In summary, the time complexity of Algorithm 1 is $O(m) + O(m) + O(n) + O(2n)$. For most real-world networks, $m > n$. Therefore, the final time complexity of Algorithm 1 is $O(m)$.

In Algorithm 2, steps 1−5 calculate the micro-community closeness, the time complexity is $O(K^2)$. Steps 6−11 detect the hierarchical community structure with Prim algorithm, a typical MST construction method. In step 6, the K micro-communities form the first level of the dendrogram. Steps 7−11 merge two micro-communities with largest micro-community closeness to form the higher level community structure at each iteration. The time complexity of steps 6−11 is $O(K \log M)$, where M is the number of edges between micro-communities. In summary, the complexity of Algorithm 2 is $O(K^2) + O(K \log M) = O(K^2)$.

The above analysis shows that the total complexity of our proposed method is $O(m) + O(K^2)$. For most real-world networks, $K << m$. Therefore, the total time complexity of our proposed method is $O(m)$.

The following Table 1 shows the characteristic and time complexity of different methods, where m is the number of edges, n is the number of vertices and k is the depth of the dendrogram. Obviously, our proposed method is more efficient.

4. Experiments and Results

4.1 Datasets, Baseline Algorithms and Evaluation Metrics

In order to evaluate the performance of our proposed method (named mMST in this paper), we compare it with three typical hierarchical community detection methods. Artificial networks includes: (1) general graphs come from LFR benchmark generator [33] which can produce the required networks with implanted communities, and (2) a hierarchical graph provided by Arenas et al. [34], which is a simple extension of Girvan and Newman benchmark [35]. It is true that, the LFR generator may not produce the exactly same artificial networks in each run even with the same parameters. In order to obtain relatively accurate results, we use the average of 50 runs as the final results.

1) Datasets. Two types of networks are used in the experiments: artificial networks and real-world networks. Ta-

Table 2 Real-world networks used in experiments

Network	n	m	Descriptions
Karate [36]	34	78	A social network about a karate club
Book [37]	105	441	A network of books about US politics
Football [32]	115	613	Network of America football games
Call1 [38]	369	525	The first snapshot of VAST 2008
Yeast [39]	2361	6646	Network of yeast cells
Face book [40]	4039	88234	Social circles from Face book (anonymized)
Power [41]	4941	6594	Network of the Western States Power Grid
Ca-GrQc [42]	5242	14490	Collaborative networks of Arxiv General Relativity
Ca-HepTh [42]	9877	25986	Collaborative networks of Arxiv High Energy Physics Theory
Ca-HepPh [42]	12008	118505	Collaborative networks of Arxiv High Energy Physics

ble 2 shows the 10 real-world networks used in experiments, n, m denote the number of vertices and edges, respectively.

2) Baseline algorithms. Three representative hierarchical community detection methods are selected as the baseline algorithms, including FN [8], CNM [9] and ANA [31]. Both FN [8] and CNM [9] are typical hierarchical community detection methods based on modularity optimization. They regard each isolated vertex as an initial community and merge the two communities with maximal modularity increment in each iteration until the entire network is merged into one. ANA [31] is a community detection method that constructs the minimum spanning tree from single vertices of networks and then detects the community structure by removing edges of the minimum spanning tree through the pre-defined rules. The time complexity of the FN, CNM and ANA algorithms are shown in Table 1.

3) Evaluation metrics. Three metrics are used to evaluate the performance of different methods, including Normalized Mutual Information (NMI) [43], Modularity [35] and running time. NMI is a widely used metric for community detection that measures the similarity between the detected community structure and the ground-truth structure. The running time reveals the efficiency of different methods. The average of 50 runs is regarded as the final runing time of each method. Modularity evaluates how good the obtained community structure is, which is defined as follow:

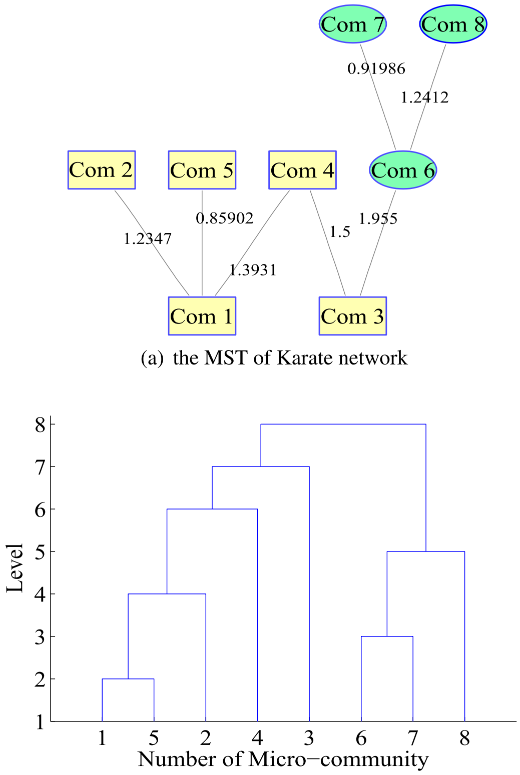
$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(c_i, c_j\right) \qquad (3)$$

In Eq. (3), A_{ij} denotes the element of the adjacent matrix of the network. If vertices i and j are connected, $A_{ij} = 1$, otherwise $A_{ij} = 0$. m refers to the edge number of the network. k_i represents the degree of vertex i, and $k_i = \sum_j (A_{ij})$. c_i is the community to which vertex i is assigned. If vertices i, j belong to the same community, i.e. $c_i = c_j$, then $\delta\left(c_i, c_j\right) = 1$, otherwise $\delta\left(c_i, c_j\right) = 0$.

4.2 Performance on Real-World Networks

Firstly, our proposed method are applied on three small scale networks, including Karate network [36], Book network [37] and Football network [32]. These three networks have well-known community structure.

Figure 2 (a) shows the MST of Karate network constructed by our proposed method. Each vertex represents a generated micro-community and the edge weight between vertices denotes the reciprocal of the micro-community closeness. These 8 micro-communities will be merged one by one, according to the edge weight. The generated dendrogram is shown in Fig. 2 (b) The 8 leaves of the dendrogram correspond to the 8 micro-communities of the Karate network. Each merging of the micro-communities will form a new hierarchy. We select the hierarchy with maximum modularity as the final community structure. The experimental results show that the 7^{th} level has the maximal modularity, therefore, the best partition is 2 communities for Karate network, as shown in Fig. 2 (a). This partition result is in accordance with the ground truth community structure of Karate network.



(a) the MST of Karate network

(b) the dendrogram of Karate network

Fig. 2 MST and corresponding dendrogram of Karate network produced by mMST

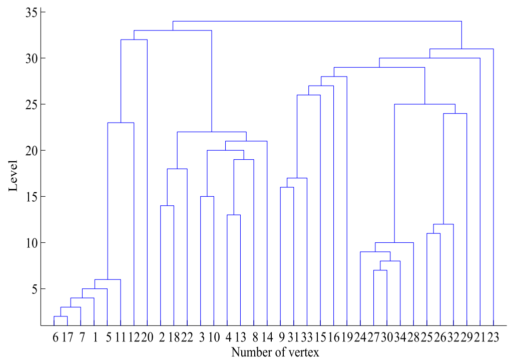
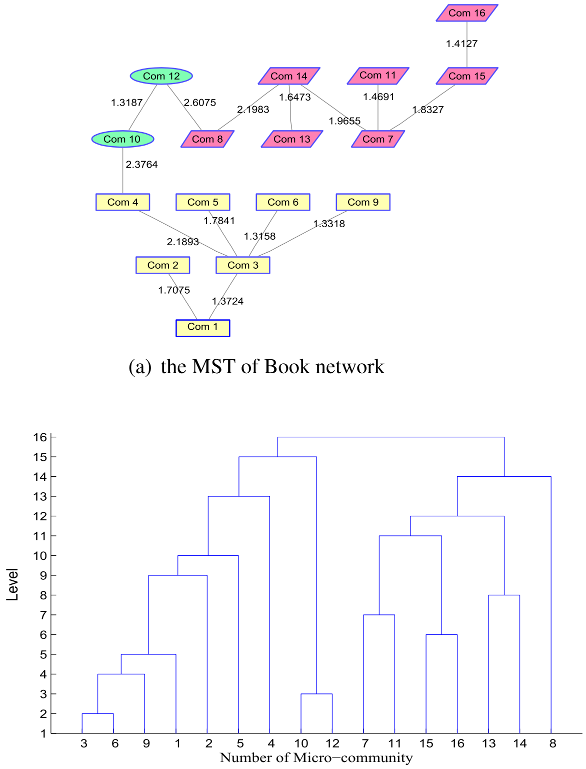


Fig. 3 Dendrogram of Karate network produced by FN algorithm.



(a) the MST of Book network

(b) the dendrogram of Book network

Fig. 4 MST and corresponding dendrogram of Book network produced by mMST

Figure 3 shows the dendrogram of Karate network generated by FN method, and the lowest level corresponds to the 34 vertices of the Karate network. The FN method treats each network vertex as an independent community and merges them one by one. We also select the hierarchy with maximum modularity as the final community structure. The experimental results show that the FN method also divides the Karate network into 2 communities, matching the real community structure. However, the dendrogram produced by FN method is more complex than that produced by our proposed method. This is because our proposed method takes each micro-community as an initial cluster while the FN method treats each individual vertex as an initial cluster. Since the number of individual vertex is much larger than that of the micro-community, the iterations of the FN method are much larger than that of the proposed method, resulting in a more complex dendrogram. Obviously, the dendrogram obtained by our method is more simple and meaningful for understanding the hierarchical community structure.

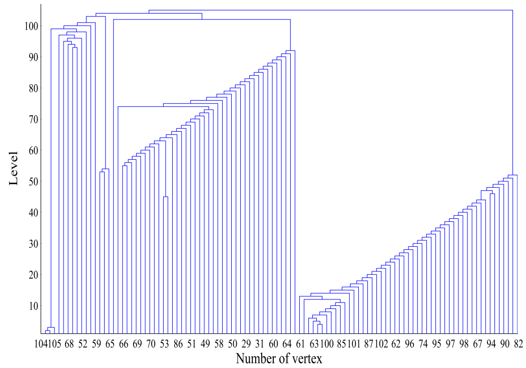
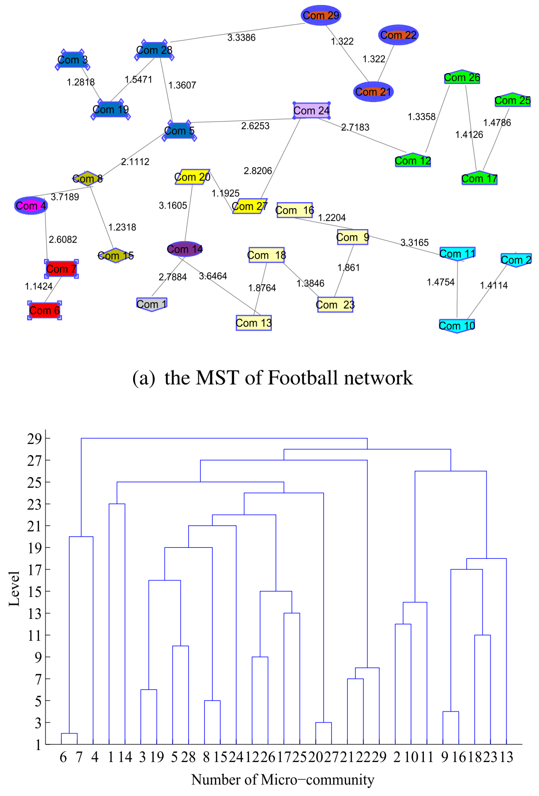


Fig. 5 Dendrogram of Book network produced by FN algorithm.



(a) the MST of Football network

(b) the dendrogram of Football network

Fig. 6 MST and corresponding dendrogram of Football network produced by mMST

Figure 4 shows the MST and corresponding dendrogram of Book network produced by our proposed method. As shown in Fig. 4 (a), the MST contains 16 vertices and each vertex corresponds to a micro-community. These 16 micro-communities will be merged one by one, forming the

Table 3 Performance of different methods on three real-world networks with ground-truth community structure

	Methods	Karate	Book	Football
Ground-truth Community number		2	3	12
Identified community number	mMST	**2**	**3**	**12**
	ANA	3	8	13
	FN	3	4	6
	CNM	3	4	7
NMI	mMST	**0.973**	**0.972**	0.853
	ANA	0.64	0.474	**0.887**
	FN	**0.973**	0.637	0.439
	CNM	0.777	0.621	0.568
running time	mMST	**0.039**	**0.305**	**1.674**
	ANA	1.677	6.298	4.223
	FN	0.469	10.084	12.723
	CNM	1.053	6.846	7.952

Table 4 Performance of different methods on seven real-world networks without ground-truth community structure

Network	mMST		FN		ANA		CNM	
	Q	Running time	Q	Running time	Q	Running time	Q	Running time
Call1	**0.596**	**3.4566**	0.568	19.246	0.112	12.551	0.552	7.683
Yeast	0.364	**118.741**	0.329	251.829	0.059	245.383	**0.365**	289.417
Face book	0.602	**692.525**	0.6	1990.932	0.106	905.512	**0.612**	1457.833
Power	0.679	**1211.522**	**0.681**	3768.249	0.129	1454.188	0.606	1977.331
Ca-GrQc	**0.537**	**1428.961**	0.513	4224.358	0.091	1678.254	0.499	2483.418
Ca-HepTh	**0.437**	**3205.694**	0.416	9359.595	0.257	5147.332	0.422	7641.845
Ca-HepPh	**0.381**	**5894.153**	**0.381**	15208.91	0.241	8735.225	0.351	12843.81

corresponding dendrogram (Fig. 4 (b)). The 16 leaves of the dendrogram correspond to the 16 micro-communities. Each merging of the micro-communities will form a new hierarchy. We also select the hierarchy with maximal modularity as the optimal partition. The experimental results show that the 14th level has the maximal modularity, which reveals a 3-communities structure, as shown in Fig. 4 (a). This partition result accurately matches the ground truth community structure of Book network.

Figure 5 shows the dendrogram of Book network produced by FN method and the lowest level refers to the 115 vertices of Book network. The optimal partition of FN method divides Book network into 4 communities, deviating from the well-known structure of this network. Furthermore, compared with Fig. 4 (b), the Fig. 5 is very complex with large depth.

Figure 6 is the MST and corresponding dendrogram of Football network generated by our proposed method. Figure 6 (a) shows 29 identified micro-communities. These 29 micro-communities will be merged one by one, forming the corresponding dendrogram (Fig. 6 (b)). The optimal partition, i.e. the 18th hierarchy with maximal modularity, divides Football network into 12 communities, in accordance with the ground truth community structure of Football network.

The above results reveal that our proposed method can accurately identify the optimal community structure. Furthermore, the obtained dendrogram is simple and intuitive, which is meaningful to analyse the hierarchical community structure of social networks.

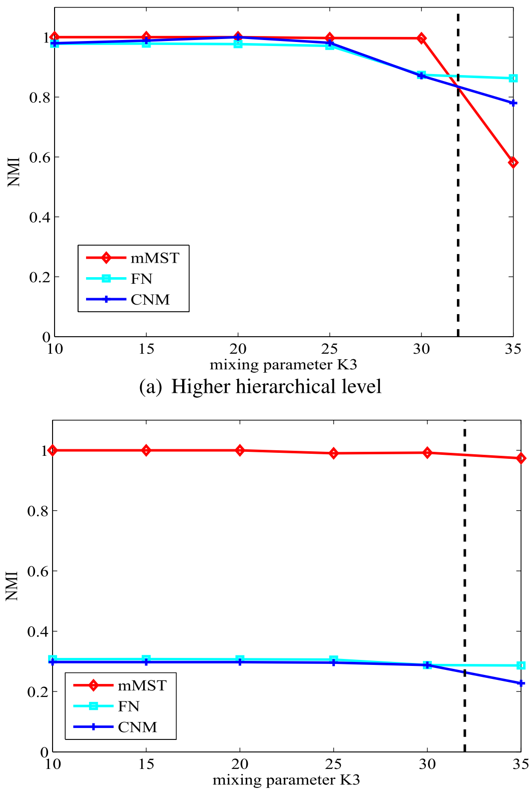
Secondly, we select the dendrogram level that maximizes the modularity as the optimal partition of different methods and further compare the performance of our mMST with that of others. Table 3 shows the corresponding results. Compared with other three methods, our proposed mMST method shows good NMI performance and identifies the community number exactly. In addition, our method can obtain the final results with very high efficiency.

Thirdly, our proposed method are applied on other seven real-world networks, including Call1 [38], Yeast [39], Face book [40], Power [41], Ca-GrQc [42], Ca-HepTh [42] and Ca-HepPh [42]. Since the ground-truth community structures of these seven networks are not available, we adopt the modularity Q to evaluate the quality of detected communities and the running time to evaluate the efficiency of different methods. We also select the dendrogram level that maximizes the modularity as the optimal partition of different methods. Table 4 shows the corresponding results. Compared with other three methods, our proposed mMST method exists good modularity Q and efficiency performance in hierarchical community partition.

4.3 Performance on Artificial Networks

4.3.1 Hierarchical Artificial Network

The hierarchical artificial graph with built-in hierarchical community structure is provided by Arenas et al. [34], which is a simple extension of Girvan and Newman benchmark [35]. This hierarchical benchmark contains 512



(a) Higher hierarchical level

(b) Lower hierarchical level

Fig. 7 NMI of different methods on the hierarchical artificial benchmark

vertices, arranged into 16 groups with 32 vertices each. The 16 groups are further ordered into 4 super-groups. In addition, each vertex has a number of K_3 links with the rest of the network ("rest" refers to the vertices that are not in the same group or super-group). In this way, two hierarchical levels emerge: the higher level contains 4 super-groups and each of them has 128 vertices; the lower level contains 16 communities with 32 vertices each. In general, the build-in hierarchy depends on the parameter K_3, the mixing degree of the four super-groups. We evaluate the NMI performance of different methods with the variation of mixing parameter. For each value, we build 50 realizations of the network.

Figure 7 (a) shows the corresponding results of the higher level which contains 4 super-communities. When $K_3 < 32$, the 4 super-communities can be correctly identified by three methods, and our proposed method exhibits a relatively better NMI performance. After that, the links outside the super-community is bigger than that within the super-community, and the 4 super-communities mixes well with each other. Thus, the community structure of the high level becomes unclear, leading to the degradation of NMI performance. Figure 7 (b) shows the corresponding results of the lower level which contains 16 communities. Our proposed method can also obtain good performance, and the identified community structure is close to the built-in structure. FN and CNM methods cannot handle the fringe vertices reasonably, and some fringe vertices are merged into communities at the higher level, resulting in poor NMI performance.

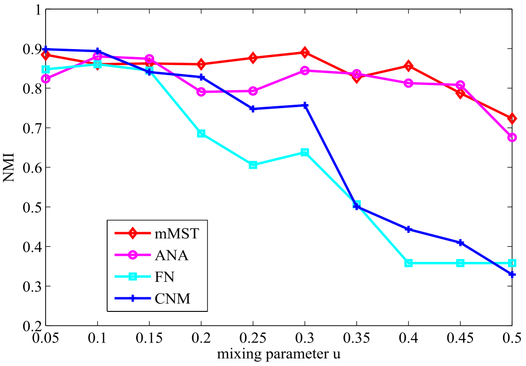


Fig. 8 NMI of different methods with the variation of the mixing parameters *u*

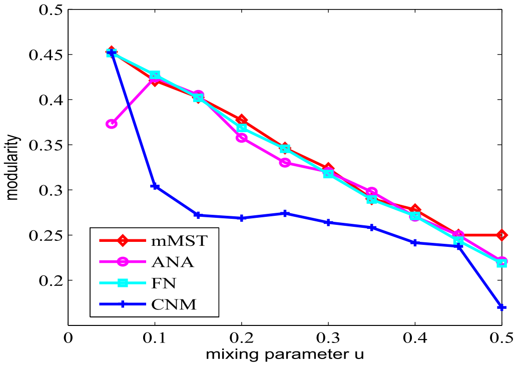


Fig. 9 Modularity of different methods with the variation of the mixing parameters *u*

4.3.2 LFR Artificial Networks

The LFR benchmark generator can be defined as follows:

$$LFR = (n, k, k_{\max}, C_{\min}, C_{\max}, t_1, t_2, u)$$

where, n denotes the number of vertices, i.e. network scale. k, $k_{\max}$ refer to the average degree and maximum degree, respectively. $C_{\min}$, $C_{\max}$ represent the minimum and maximum size of communities, respectively. t_1, t_2 are the exponents of the power-law distributions of vertex degree and community size, respectively. u denotes a mixing parameter used to regulate the quality of the community structure to be generated.

Firstly, we analyze the influences of the mixing parameter u on partition performance. Figure 8, Fig. 9 show the NMI and modularity of different methods with the variation of the mixing parameter u, respectively. The other key parameters are set as follows: $n = 5000$, $k = 15$, $k_{\max} = 50$, $C_{\min} = 20$, $C_{\max} = 50$, $t_1 = 2$, $t_2 = 1$. With the increasing of u, the community structure of the generated artificial network becomes unclear, leading to the degradation of NMI performance (Fig. 8) and the deterioration of modularity value (Fig. 9). Compared with other three methods, our mMST method can obtain the best NMI and modularity performance for the most value of the mixing parameter u.

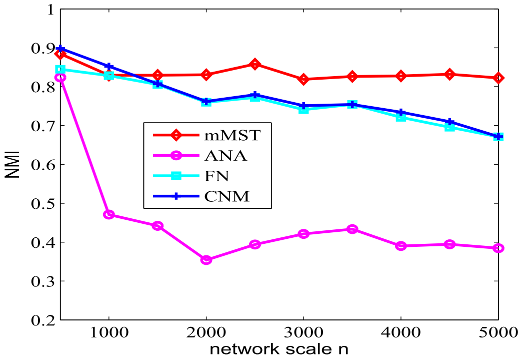


Fig. 10 NMI of different methods with the variation of the network scale n

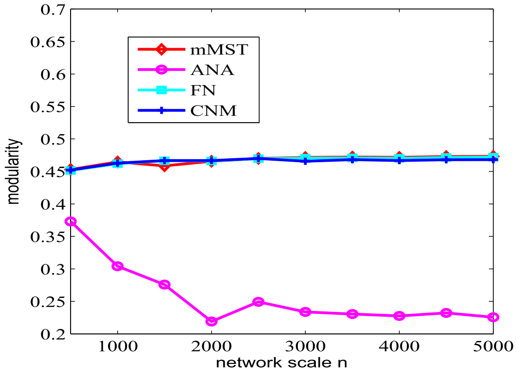


Fig. 11 Modularity of different methods with the variation of the network scale n

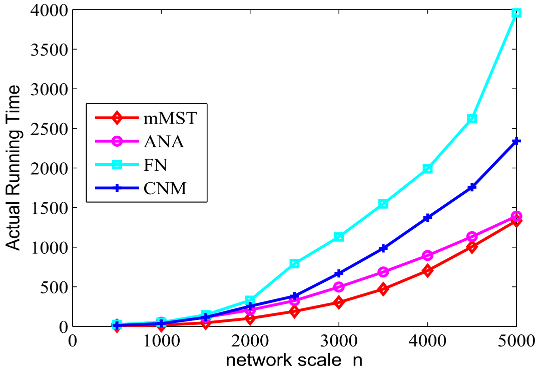


Fig. 12 Running time of different methods with the variation of the network scale n

Secondly, we analyze the influences of the network scale n on partition performance. Figure 10, Fig. 11 show the NMI and modularity of different methods with the variation of the network scale n, respectively. The other key parameters are set as follows: $k = 15$, $k_{\max} = 50$, $C_{\min} = 20$, $C_{\max} = 50$, $u = 0.05$, $t_1 = 2$, $t_2 = 1$. With the increasing of the network scale n, our proposed method shows the best NMI performance. mMST also exhibits competitive modularity performance as that of FN and CNM. We further record the running time of different methods with the variation the network scale n, and Fig. 12 shows the corresponding results. Each point denotes the average of 50 times runs. Compared with other three methods, the curve of our proposed method rises more slowly with the increasing of n, indicating a high partition efficiency.

5. Conclusion

Many social networks exhibit a natural hierarchical community structure. Most traditional methods need some predefined parameters, and the obtained dendrograms are extremely complex to understand and analyze. In the paper, we propose a novel hierarchical community detection method based on micro-community and minimum spanning tree. The proposed method calculates the link strength of adjacent vertices and then generates micro-communities. Minimum spanning tree is constructed from these generated micro-communities by linking one by one. The linking order will reveal the nest structure of these micro-communities. We applied our proposed method on synthetic and real-world networks. Compared with other state-of-the-art methods, our proposed method exhibits good accuracy and efficiency performance. In addition, our proposed method does not require any pre-defined parameters, and the output dendrogram is very intuitively for understanding and analyzing.

Our proposed method mainly concentrates on the undirected and unweighted networks. In the future, we will further investigate directed networks and weighted networks. What's more, Hadoop or Spark will be adopted for parallel processing to achieve better performance, especially for large scale networks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.61876186, No.71774159).

References

[1] Z. Wang, J. Xi, Y. Xing, and Z. Hu, "Community number estimation for community detection in complex networks," Journal of Information Science and Engineering, vol.33, no.5, pp.1323–1341, 2017.

[2] W. Zhi-Xiao, L. Ze-chao, D. Xiao-fang, and T. Jin-hui, "Overlapping community detection based on node location analysis," Knowledge-Based Systems, vol.105, pp.225–235, 2016.

[3] W. Wang and W.N. Street, "Finding hierarchical communities in complex networks using influence-guided label propagation," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp.547–556, IEEE, 2015.

[4] S. Zhao, C. Yu, and Y. Zhang, "Hierarchical community detection based on multi degrees of distance space and submodularity optimization," Chinese National Conference on Social Media Processing, pp.343–354, Springer, 2017.

[5] G. Cai, R. Wang, and G. Liu, "Hierarchical overlapping community discovery algorithm based on node purity," International Conference on Intelligent Information Processing, pp.248–257, Springer, 2012.

[6] D. Guohui, S. Huimin, F. Chunlong, and S. Yan, "Community detection algorithm of the large-scale complex networks based on random walk," International Conference on Web-Age Information Management, pp.269–282, Springer, 2016.

[7] M. Shirzad and M.R. Feizi-Derakhshi, "Hierarchical community detection in social networks using spectral method," International Journal of Computer Science and Information Security, vol.14, no.8, pp.51–59, 2016.

[8] M.E.J. Newman, "Fast algorithm for detecting community structure in networks," Physical review E, vol.69, no.6, p.066133, 2004.

[9] A. Clauset, M.E.J. Newman, and C. Moore, "Finding community structure in very large networks," Physical review E, vol.70, no.6, p.066111, 2004.

[10] A.M. Fiscarelli, A. Beliakov, S. Konchenko, and P. Bouvry, "A degenerate agglomerative hierarchical clustering algorithm for community detection," Asian Conference on Intelligent Information and Database Systems, pp.234–242, Springer, 2018.

[11] B. Saoud and A. Moussaoui, "A new hierarchical method to find community structure in networks," Physica A: Statistical Mechanics and its Applications, vol.495, pp.418–426, 2018.

[12] R. Franke, "Chimera: Top-down model for hierarchical, overlapping and directed cluster structures in directed and weighted complex networks," Physica A: Statistical Mechanics and its Applications, vol.461, pp.384–408, 2016.

[13] R. Toujani and J. Akaichi, "Optimal initial partitionning for high quality hybrid hierarchical community detection in social networks," 2017 4th International Conference on Control, Decision and Information Technologies (CoDIT), pp.0395–0403, IEEE, 2017.

[14] R. Toujani and J. Akaichi, "A model based metaheuristic for hybrid hierarchical community structure in social networks," World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol.11, no.6, pp.661–666, 2017.

[15] M.J. Barber, "Detecting hierarchical communities in networks: A new approach," Stochastic and Infinite Dimensional Analysis, pp.19–37, Springer, 2016.

[16] C.-C. Lin, J.-R. Kang, and J.-Y. Chen, "An integer programming approach and visual analysis for detecting hierarchical community structures in social networks," Information Sciences Informatics and Computer Science, Intelligent Systems, Applications: An International Journal, vol.299, pp.296–311, 2015.

[17] J. Huang, H. Sun, J. Han, and B. Feng, "Density-based shrinkage for revealing hierarchical and overlapping community structure in networks," Physica A: Statistical Mechanics and its Applications, vol.390, no.11, pp.2160–2171, 2011.

[18] N. Schlitter, T. Falkowski, et al., "Dengraph-ho: Density-based hierarchical community detection for explorative visual network analysis," International Conference on Innovative Techniques and Applications of Artificial Intelligence, pp.283–296, Springer, 2011.

[19] K. Subramani, A. Velkov, I. Ntoutsi, P. Kroger, and H.-P. Kriegel, "Density-based community detection in social networks," 2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application, pp.1–8, IEEE, 2011.

[20] X. Qi, W. Tang, Y. Wu, G. Guo, E. Fuller, and C.-Q. Zhang, "Optimal local community detection in social networks based on density drop of subgraphs," Pattern Recognition Letters, vol.36, no.1, pp.46–53, 2014.

[21] P. Kim and S. Kim, "Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering," Physica A: Statistical Mechanics and its Applications, vol.417, pp.46–56, 2015.

[22] L. Zhou, K. Lü, P. Yang, L. Wang, and B. Kong, "An approach for overlapping and hierarchical community detection in social networks based on coalition formation game theory," Expert Systems with Applications, vol.42, no.24, pp.9634–9646, 2015.

[23] L. Zhou, C. Cheng, K. Lü, and H. Chen, "Using coalitional games to detect communities in social networks," International Conference on Web-Age Information Management, pp.326–331, Springer, 2013.

[24] W. Zhang, F. Kong, L. Yang, Y. Chen, and M. Zhang, "Hierarchical community detection based on partial matrix convergence using random walks," Tsinghua Science and Technology, vol.23, no.1, pp.35–46, 2018.

[25] J. Qiu and Z. Lin, "D-hocs: an algorithm for discovering the hierarchical overlapping community structure of a social network," Journal of Intelligent Information Systems, vol.42, no.3, pp.353–370, 2014.

[26] B. Yang, J. Di, J. Liu, and D. Liu, "Hierarchical community detection with applications to real-world network analysis," Data & Knowledge Engineering, vol.83, no.4, pp.20–38, 2013.

[27] R. Du, D. Kuang, B. Drake, and H. Park, "Hierarchical community detection via rank-2 symmetric nonnegative matrix factorization," Computational social networks, vol.4, no.1, p.7, 2017.

[28] N. Gillis, D. Kuang, and H. Park, "Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization," IEEE Trans. Geosci. Remote Sens., vol.53, no.4, pp.2066–2078, 2015.

[29] R.K. Behera, S.K. Rath, and M. Jena, "Spanning tree based community detection using min-max modularity," Procedia Computer Science, vol.93, pp.1070–1076, 2016.

[30] B. Saoud and A. Moussaoui, "Community detection in networks based on minimum spanning tree and modularity," Physica A: Statistical Mechanics and its Applications, vol.460, pp.230–234, 2016.

[31] K. Asmi, D. Lotfi, and M. El Marraki, "A novel approach based on the minimum spanning tree to discover communities in social networks," 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM), pp.286–290, IEEE, 2016.

[32] M. Girvan and M.E.J. Newman, "Community structure in social and biological networks," Proc. national academy of sciences, vol.99, no.12, pp.7821–7826, 2002.

[33] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical review E, vol.78, no.4, p.046110, 2008.

[34] A. Arenas, A. Díaz-Guilera, and C.J. Pérez-Vicente, "Synchronization reveals topological scales in complex networks," Physical review letters, vol.96, no.11, p.114102, 2006.

[35] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol.69, no.2, p.026113, 2004.

[36] W.W. Zachary, "An information flow model for conflict and fission in small groups," Journal of anthropological research, vol.33, no.4, pp.452–473, 1977.

[37] M.E.J. Newman, "Modularity and community structure in networks," Proc. national academy of sciences, vol.103, no.23, pp.8577–8582, 2006.

[38] Y. Liu, H. Gao, X. Kang, Q. Liu, R. Wang, and Z. Qin, "Fast community discovery and its evolution tracking in time-evolving social networks," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp.13–20, IEEE, 2015.

[39] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, "Topological structure analysis of the protein–protein interaction network in budding yeast," Nucleic acids research, vol.31, no.9, pp.2443–2450, 2003.

[40] J. Leskovec and J.J. Mcauley, "Learning to discover social circles in ego networks," Advances in Neural Information Processing Systems, pp.539–547, 2012.

[41] D.J. Watts and S.H. Strogatz, "Collective dynamics of 'small-world' networks," nature, vol.393, no.6684, pp.440–442, 1998.

[42] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," ACM Transactions on Knowledge Discovery from Data (TKDD), vol.1, no.1, p.2, 2007.

[43] C. Giatsidis, F.D. Malliaros, D.M. Thilikos, and M. Vazirgiannis, "Corecluster: A degeneracy based graph clustering framework," AAAI, pp.44–50, 2014.

Zhixiao Wang received his Ph.D. degree in the Department of Computer Science and Engineering at Tongji University in 2011. Currently, he served as an associate professor at China University of Mining and Technology. His research interests include information diffusion, social network analysis.

Mengnan Hou is a M.E. candidate at School of Computer Science and Technology, China University of Mining Technology. Her main research interests include data mining and hierarchical community detection.

Guan Yuan received his B.E., M.E., Ph.D degrees in Computer Science and Technology from China University of Mining and Technology in 2004, 2009 and 2012, respectively. Currently, he served as an associate professor at China University of Mining and Technology. His research interests include data mining, especially moving object data mining.

Jing He is a M.E. candidate at School of Computer Science and Technology, China University of Mining Technology. Her main research interests include dynamic community detection and community evolution analysis.

Jingjing Cui graduated from School of Computer Science and Technology, China University of Mining and Technology in 2010. He is the key account director of Baidu Cloud, Baidu Online Network Technology (Beijing) Co., Ltd. His research interests include cloud computing, social network analysis.

Mingjun Zhu is a M.E. candidate at School of Computer Science and Technology, China University of Mining Technology. His main research interests include trajectory data analysis and mining.