Community Detection using Closeness Similarity based on Common Neighbor Node Clustering Entropy

Wanchang Jiang[1], Xiaoxi Zhang[1], and Weihua Zhu[2*]
[1] School of Computer Science, Northeast Electric Power University
Jilin 132012, China
[e-mail: jwchang84@163.com, zxxyx0711@163.com]
[2] Department of Information Technology, Jilin Technology College of Electronic Information
Jilin 132121, China
[e-mail: 317010530@qq.com]
[*]Corresponding author: Weihua Zhu

Received March 17, 2022; revised May 27, 2022; accepted August 1, 2022; published August 31, 2022

Abstract

In order to efficiently detect community structure in complex networks, community detection algorithms can be designed from the perspective of node similarity. However, the appropriate parameters should be chosen to achieve community division, furthermore, these existing algorithms based on the similarity of common neighbors have low discrimination between node pairs. To solve the above problems, a noval community detection algorithm using closeness similarity based on common neighbor node clustering entropy is proposed, shorted as CSCDA. Firstly, to improve detection accuracy, common neighbors and clustering coefficient are combined in the form of entropy, then a new closeness similarity measure is proposed. Through the designed similarity measure, the closeness similar node set of each node can be further accurately identified. Secondly, to reduce the randomness of the community detection result, based on the closeness similar node set, the node leadership is used to determine the most closeness similar first-order neighbor node for merging to create the initial communities. Thirdly, for the difficult problem of parameter selection in existing algorithms, the merging of two levels is used to iteratively detect the final communities with the idea of modularity optimization. Finally, experiments show that the normalized mutual information values are increased by an average of 8.06% and 5.94% on two scales of synthetic networks and real-world networks with real communities, and modularity is increased by an average of 0.80% on the real-world networks without real communities.

Keywords: Complex Network, Community Detection, Node Similarity, Closeness, Common Neighbor Node

This work is supported in part by the Research Project of the Education Department of Jilin Province (JJKH20220111KJ)

http://doi.org/10.3837/tiis.2022.08.007 **ISSN : 1976-7277**

1. Introduction

Many complex systems in the real world can be abstracted into networks, such as social networks [1], gene regulatory networks [2], transportation networks [3], power networks [4], etc. Nodes with similar attributes in the network are often easy to form groups, which are manifested in community or module structure. Community detection can help us understand the function of complex networks and predict the behaviors of complex networks. At present, community detection algorithms are widely used in social network recommendation [5], biological protein integration [6], network public opinion analysis [7] and so on.

According to the algorithm purpose, community detection algorithms can generally be classified into two categories, that is, non-overlapping and overlapping community detection algorithms. According to the algorithm idea, community detection algorithms can be classified into the algorithm based on graph segmentation [8], the algorithm based on label propagation [9], and the algorithm based on hierarchical clustering [10]. Community detection algorithms have been deeply researched, and a detailed review of these algorithms will be described in Section 2. Among them, the hierarchical clustering algorithm can detect communities by splitting or condensing based on the similarity or strength of the connections between nodes, which has the advantages of simpleness and efficiency. However, there are still problems in the construction of node similarity and the selection of algorithm parameters. Therefore, in view of the low accuracy and the difficulty of parameter setting in similarity-based hierarchical clustering algorithms, a noval community detection algorithm is proposed, which can effectively measure the similarity between node pairs to improve the accuracy of the detection result, and realize the stable community detection without parameters through the merging of two levels. The major contributions of this paper are as follows.

1) To obtain the differential of common neighbor nodes of the node pair when calculating the node similarity, we design a closeness similarity measure by the defined common neighbor node clustering entropy. Considering the closeness information provided by a common neighbor node to its all first-order neighbor nodes, the node similarity can be effectively calculated.
2) To detect communities accurately and stably without parameters, we propose a closeness similarity-based community detection algorithm. By using the designed closeness similarity measure and node leadership, the initial communities are formed. Then, based on the idea of modularity optimization, the final communities are detected by the merging of two levels.
3) To verify the effectiveness and robustness of the proposed algorithm, experiments are carried out on two scales of synthetic networks and real-world networks which are divided into disassortative and assortative networks. Compared with the other three algorithms, the proposed algorithm can detect high-quality communities in complex networks.

The structure of the paper is organized as follows: Section 2 introduces the related work. In Section 3, we propose a noval community detection algorithm using closeness similarity based on common neighbor node clustering entropy. In Section 4, the performance of the proposed algorithm is demonstrated by using normalized mutual information and modularity on synthetic networks and real-world networks. Finally, some conclusions and future works are given in Section 5.

2. Related Work

For the research of community detection, scholars have proposed many different methods, which mainly include the algorithm based on graph segmentation, the algorithm based on label propagation, and the algorithm based on hierarchical clustering.

Graph segmentation algorithm usually divides nodes into certain numbers and sizes of communities, so that the community inside has edges as many as possible [8, 11]. However, the number of dividing communities must be ascertained ahead of time and it cannot guarantee the optimal result.

With the expansion of the network scale, the label propagation algorithm (LPA) is proposed to reduce the time complexity [9, 12]. It mainly utilizes the neighbor information of each node to ascertain its label without knowing the community structure in advance. As a classical label propagation algorithm, LPA iteratively designates a single label for each node and examines the neighbor nodes of each node until each node has the same label with most of its neighbor nodes [13]. However, LPA is very sensitive to label update rules and its result is random and unstable. To overcome this shortcoming, many scholars have done a lot of attempts. In 2020, Zhang et al. [14] designed a new label propagation mechanism to cope with instability, which were influenced by human society and radar transmission. In addition, a parallel label propagation algorithm based on weight and random walk (WRWPLPA) [15] is proposed. In the process of label propagation, the stability of the algorithm is greatly improved by calculating the weights. It can be seen that many scholars have further improved the LPA algorithm to solve the problem of instability. But for large sparse networks, the LPA algorithm may lead to the emergence of giant communities.

Since the number and size of clustering do not need to be known in advance and giant communities are not generated, the hierarchical clustering algorithm has attracted much attention from many researchers. And community detection from the perspective of similarity is also an important research method. Splitting and agglomeration are two well-known hierarchical clustering strategies [16]. Split-based methods first treat the entire network as a whole and then divide it into groups according to the predefined rules. Zarandi et al. [17] arbitrarily deleted some edges in the network according to the similarity between edges, resulting in some subgraphs as the main communities, and then some subgraphs were merged to get optimal communities. In the pairing, splitting and aggregating algorithm (PSA) [18], the whole network is split into several similar node sets as an essential step to detect the final communities.

In contrast, agglomerative methods first create many initial groups and then merge them according to different similarity calculation methods. Wang et al. [10] used the Jaccard index and degree clustering information as a local similarity measure to extract communities. However, the problems of randomness and uncertainty may arise when selecting the most similar node. Based on this, Zhang et al. [19] provided a multi-level similarity calculation approach, and a new community detection model is designed. And Liu et al. [20] designed the local community detection algorithm using fuzzy similar relationships. To effectively detect communities with complex structures, HCLORE [21] obtains initial communities by searching local kernels. However, HCLORE is prone to errors in assigning nodes to initial communities due to the difficulty of finding suitable nuclei in sparse communities. Furthermore, since the existing algorithms detect communities without considering visual understanding of communities, by defining node leadership and membership, the simplified tree-based community detection approach (STCD) [22] is proposed. The quantity of common neighbor nodes is used to estimate node membership, but the difference of common neighbors is ignored in this way. The same problem also exists in the local node similarity algorithm

(NSA) [23], besides, STCD and NSA all need to be set suitable parameters to obtain the optimal partition result.

Thus, in similarity-based hierarchical clustering community detection algorithms, the number of common neighbor nodes is usually used to measure the similarity between nodes. However, many node pairs cannot be distinguished due to the same similarity value and the selection of algorithm parameters is difficult, resulting in unstable and inaccurate community detection results. To this end, a noval community detection algorithm is proposed. On the basis of common neighbors, we consider the difference between different common neighbor nodes in the clustering coefficient and reflect it in the form of entropy, then we design a closeness similarity measure to effectively distinguish node pairs. The initial communities are constructed by using this closeness similarity measure, and the final community detection is realized by the merging of two levels without setting parameters.

3. The Proposed Community Detection Algorithm

Common neighbors are used for calculating node similarity in similarity-based hierarchical clustering community detection algorithms [10, 22, 23]. However, the quantity and degree values of common neighbors among different node pairs are usually the same in complex networks. At this time, when the initial communities are formed, the randomness and uncertainty of node selection will reduce the accuracy of community detection.

Therefore, a noval similarity-based community detection algorithm is proposed. Firstly, for calculating the similarity of one node pair based on common neighbor nodes, the common neighbor node clustering entropy is defined to design a closeness similarity measure. Then, with the idea of modularity optimization, the community detection algorithm is proposed to detect communities using the closeness similarity measure.

3.1 Closeness Similarity Measure

For an undirected and unweighted network $G=(V,E)$, $V=\{v_i \mid i=1,2,...,n\}$ is the set of n nodes and $E=\{e_{ij} \mid e_{ij}=(v_i,v_j),\ v_i \in V,\ v_j \in V \text{ and } i \neq j\}$ is the set of m edges. For any two nodes v_i and v_j in G, if an edge e_{ij} exists between v_i and v_j, then v_i and v_j are called node pair $<v_i,v_j>$, that is, v_i is a first-order neighbor node of v_j and v_j is also a first-order neighbor node of v_i. $N(v_i)=\{v_j \mid e_{ij} \in E\}$ is the set of all first-order neighbor nodes of v_i. $d(v_i)=|N(v_i)|$ is the degree of v_i. If v_z is a first-order neighbor node of both v_i and v_j, then v_z is called the common neighbor node of the node pair $<v_i,v_j>$. The set of all common neighbor nodes of the node pair $<v_i,v_j>$ can be defined as $CN(v_i,v_j)$:

$$CN(v_i,v_j)=\{v_z \mid v_z \in N(v_i) \cap N(v_j) \text{ and } e_{ij} \in E\} \tag{1}$$

There may be a good deal of nodes with identical degree values in G, but their clustering coefficients may be different. The clustering degree between $v_z \in CN(v_i,v_j)$ and its all first-order neighbor nodes can be described by the clustering coefficient CC_z of v_z. CC_z is

shown as follows:

$$CC_z = \frac{2\left|E(v_z)\right|}{d(v_z)(d(v_z)-1)} \tag{2}$$

where $E(v_z) = \left\{ e_{pq} \left| e_{pq} \in E, v_p \in N(v_z), v_q \in N(v_z) \text{ and } p \neq q \right. \right\}$ is the set of all connected edges among first-order neighbor nodes of v_z.

Based on the entropy model, the clustering coefficient is used to define common neighbor node clustering entropy.

Definition 1. (Common neighbor node clustering entropy): For each common neighbor node $v_z \in CN(v_i, v_j)$, the clustering coefficient CC_z of v_z can be used to evaluate the closeness information provided by v_z to its all first-order neighbor nodes through the entropy model, so the common neighbor node clustering entropy is defined as CE_z, which is calculated as follows:

$$CE_z = -CC_z \times \log_2 CC_z \tag{3}$$

the larger the clustering coefficient CC_z of v_z is, the closer the relationship between v_z and its all first-order neighbor nodes is, therefore, the smaller closeness information can be contributed by v_z to node pair $< v_i, v_j >$.

In the similarity calculation of each node pair, the closeness of the node pair and its corresponding each common neighbor node is considered, so common neighbor node clustering entropy-based closeness similarity is defined and calculated.

Definition 2. (Common neighbor node clustering entropy-based closeness similarity): The common neighbor node clustering entropy CE_z represents the amount of information brought by $v_z \in CN(v_i, v_j)$, namely, the closeness of v_z to its all first-order neighbor nodes. The closeness similarity of node pair $< v_i, v_j >$ can be calculated by common neighbor node clustering entropy between v_i and v_j, so the common neighbor node clustering entropy-based closeness similarity is defined as $sim(v_i, v_j)$, which is calculated as follows:

$$sim(v_i, v_j) = \begin{cases} \dfrac{\sum\limits_{v_z \in CN(v_i, v_j)} CE_z + 1}{d(v_i)} & \left|CN(v_i, v_j)\right| \geq 1 \\ \dfrac{1}{d(v_i)} & \left|CN(v_i, v_j)\right| = 0 \end{cases} \tag{4}$$

when the common neighbor nodes set $\left|CN\left(v_i,v_j\right)\right|=0$, the closeness similarity of the node pair $<v_i,v_j>$ is calculated by the degree $d(v_i)$ of v_i. The '1' in the molecule prevents the closeness similarity of node pair $<v_i,v_j>$ from being 0 due to no common neighbor nodes.

On the basis of the closeness similarity, the closeness similar node set of v_i can be defined as $CSNS(v_i)$:

$$CSNS(v_i)=\left\{v_j\left|\max_{v_j\in N(v_i)} sim\left(v_i,v_j\right)\right.\right\} \tag{5}$$

Furthermore, leadership[22] can evaluate the attractiveness of v_i to its first-order neighbor nodes, which depends on the quantity of common neighbor nodes of v_i and its neighbor nodes whose degree values are less than $d(v_i)$, so node leadership L_i can be calculated as follows:

$$L_i=\sum_{d(v_i)>d\left(v_j\right),v_j\in N(v_i)}\left|CN(v_i,v_j)\right| \tag{6}$$

the node v_i can only lead its first-order neighbor nodes, and the influence of these neighbor nodes is lower than that of itself. If some of first-order neighbor nodes of v_i have a higher influence than that of itself, the node v_i cannot lead these high influence neighbor nodes.

Therefore, in order to distinguish nodes in $CSNS(v_i)$, leadership L_j of the neighbor node v_j of v_i can be calculated as follows:

$$L_j=\sum_{d\left(v_j\right)>d(v_k),v_k\in N\left(v_j\right)}\left|CN(v_j,v_k)\right|\ ,\ v_j\in CSNS\left(v_i\right) \tag{7}$$

By selecting v_j with the largest leadership in $CSNS(v_i)$, the most closeness similar node of v_i can be identified as $v_j^{L_{\max}}$.

3.2 Closeness similarity-based community detection algorithm

In NSA [23], the similarity measure is used for forming initial communities, then by using the selected community metric parameter, the final communities are detected by merging the initial communities. Though the problem of community resolution limit is overcome, there are still some problems. First, in the forming initial communities, the use of the Jaccard index cannot effectively calculate node similarity. Second, in the merging communities, whether the community metric parameter is appropriate will affect the algorithm performance.

Based on the idea of two-stage community detection in NSA, by designing the closeness similarity based on common neighbor node clustering entropy and using the merging of two levels based on the idea of modularity optimization, a noval closeness similarity-based community detection algorithm (CSCDA) is proposed. The flowchart of the proposal is

illustrated in **Fig. 1**, which mainly includes two parts.

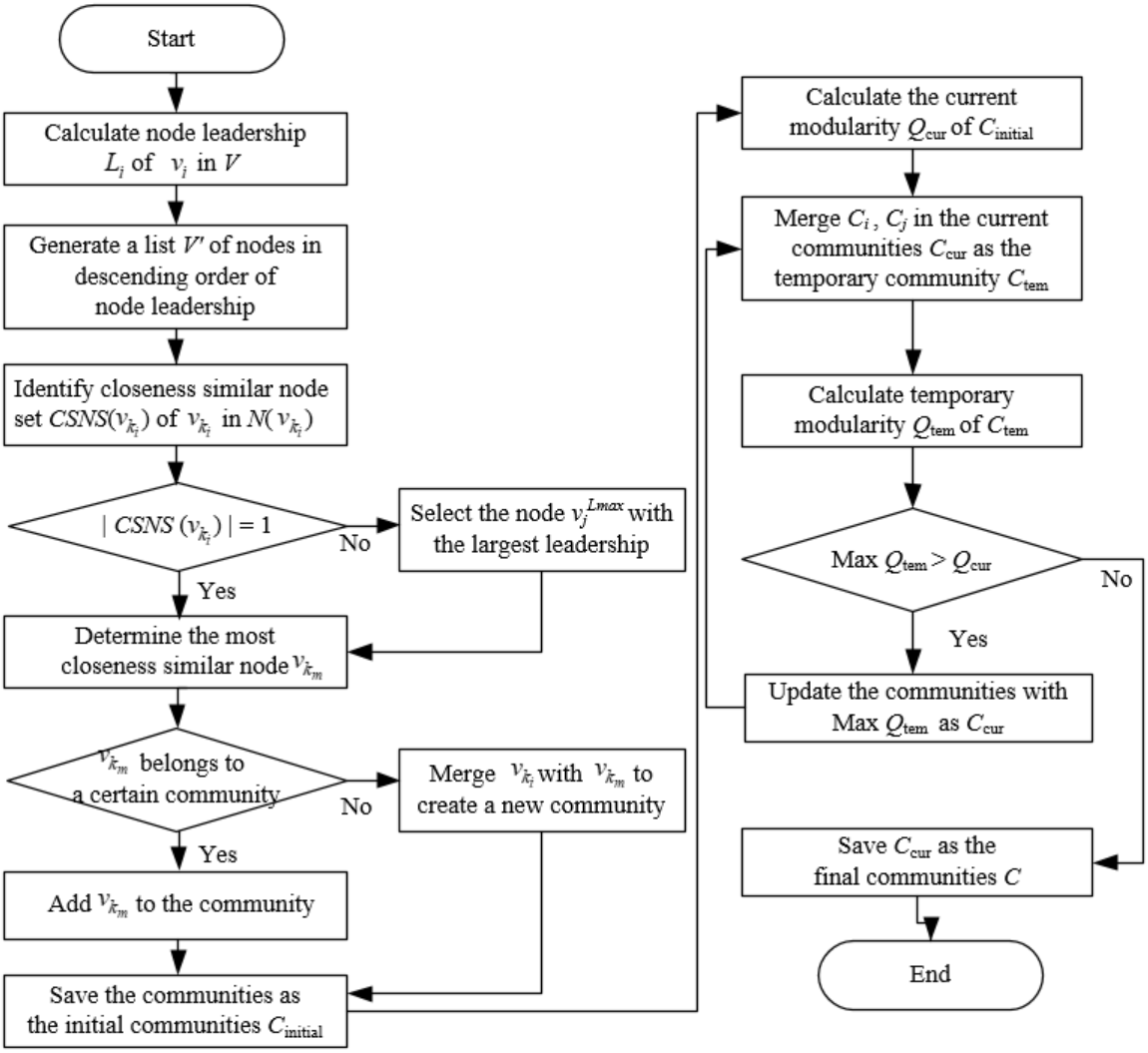


Fig. 1. The flowchart of the proposed algorithm

1) Forming initial communities: Calculating the node leadership of all nodes in the network. Processing each node from the node with the largest leadership, by using the designed closeness similarity measure, closeness similar nodes can be identified, and the node with the largest leadership is selected as the most closeness similar node. Then if the most closeness similar node already belongs to a certain community, the current node is added to the community, otherwise, the most closeness similar node will be merged with the current node to create a new community. Visiting and processing all nodes in turn until each node belongs to one community. The initial communities are formed.

2) Merging communities: The modularity of the initial communities is calculated as the current modularity. With the idea of modularity optimization, the merging of two levels is used to iteratively update communities. In the first-level merging, each two communities in the initial communities are merged as the temporary community, and temporary modularity is calculated. In the second-level merging, if the maximum temporary modularity is greater than

the current modularity, the maximum temporary modularity and its corresponding communities are updated as the current modularity and the current communities. Then the merging of two levels is repeated until the maximum temporary modularity is less than the current modularity. The final communities are detected.

The specific steps of CSCDA are as follows:

Input: Undirected and unweighted network $G=(V,E)$

Output: The detected communities C

// The first stage: Forming initial communities

1) For each $v_i \in V\ (i=1\to n)$ do
2) Calculate node leadership L_i by Formula (6);
3) Rank nodes' leadership values in descending order, denote it as $V'=\{v_{k_1},v_{k_2},...,v_{k_n}\}$;
4) Initialize $C_{\text{initial}}=\{\ \}$;
5) For each $v_{k_i} \in V'\ (i=1\to n)$ do
6) For each $v_j \in N(v_{k_i})$ do
7) Calculate $sim(v_{k_i},v_j)$ by Formula (4);
8) Get $CSNS(v_{k_i})$ by Formula (5);
9) Get $v_j^{L_{\max}}$ from $CSNS(v_{k_i})$;
10) Find $v_{k_m} \in V'$ corresponding to $v_j^{L_{\max}}$;
11) If v_{k_m} belongs to the community $C_t \in C_{\text{initial}}$ then
12) $C_t=C_t \bigcup \{v_{k_i}\}$;
13) Initialize $C_i=\{\ \}$;
14) Put C_i in C_{initial};
15) Else
16) Create a new community $C_i=\{v_{k_i},v_{k_m}\}$;
17) Initialize $C_m=\{\ \}$;
18) Put C_i, C_m in C_{initial};
19) Delete v_{k_m} from V';
20) Get $C_{\text{initial}}=\{C_1,C_2,...,C_n\}$;

// The second stage: Merging communities

21) Denote C_{initial} as C_{cur};
22) Calculate the current modularity Q_{cur};
23) For each $C_i \in C_{\text{cur}}\ (i=1\to n)$ do
24) For each $C_j \in C_{\text{cur}}\ (j=1\to n)$ do
25) If $C_i \neq C_j$ and $C_i, C_j \neq \varnothing$ then
26) $C_i=C_i \bigcup C_j$;
27) Initialize $C_j=\{\ \}$;

28) $C_{\text{tem}} = C_{\text{cur}}$;
29) Calculate the temporary modularity Q_{tem};
30) Select $Q' = \max Q_{\text{tem}}$;
31) If $Q' > Q_{\text{cur}}$ then
32) $Q_{\text{cur}} = Q'$;
33) $C_{\text{cur}} = C_{\text{tem}}$, C_{tem} is the communities with $\max Q_{\text{tem}}$, goto Step 23);
34) For each $C_i \in C_{\text{cur}} \left(i = 1 \to n\right)$ do
35) If $C_i \neq \varnothing$ then
36) Delete C_i from C_{cur};
37) Return C;

4. Experiments

4.1 Dataset description

To evaluate the performance of CSCDA, experiments are performed using synthetic networks and real-world networks, that is, synthetic networks based on Lancichinetti-Fortunato-Radicchi (LFR) benchmark [24] and real-world networks from the Konect project [25].

A. Synthetic Networks

The synthetic networks include LFR500 and LFR1000 benchmark networks, and their parameters configuration is shown in **Table 1**, where n is the number of nodes, $<d>$ and $d_{\max}$ are the average degree and maximum degree of each node, $\exp_d$ and $\exp_{com}$ are the exponents of node degree and community size according to the power-law distribution, $C_{\min}$ and $C_{\max}$ represent the minimum and maximum number of nodes contained in each community, μ represents the mixing parameter.

Table 1. Parameters configuration of LFR500 and LFR1000

Network	n	$<d>$	$d_{\max}$	$\exp_d$	$\exp_{com}$	$C_{\min}$	$C_{\max}$	μ
LFR500	500	20	50	2	1	20	100	0.1~0.9
LFR1000	1000	20	50	2	1	20	100	0.1~0.9

B. Real-World Networks

The basic information of Karate network, Risk Map network, Dolphin network, Football network, Physicians network, and Email network is shown in **Table 2**, where n and m represent the number of nodes and edges, dc and cc represent the degree correlation and the average clustering coefficient of the network, respectively.

The Karate network is the statistical information provided by the sociologist Zachary based on the relationships between members of the karate club. Club managers, coaches, and members are regarded as nodes, and their friendship is abstracted as the edge. The Risk Map network is a world map loaded in the board game, Risk. Nodes represent countries, and each edge represents the geographically adjacent relationship between two countries. The Dolphin network describes the associations among dolphin groups in New Zealand. Each node

represents the interaction between dolphin species and each edge represents the interaction between two dolphins. The Football network was collected by Newman and Girvan from 115 American college football teams. Each node represents a team and each edge represents a match between two teams. The Physicians network is a directed network, where a node represents a physician and an edge represents the communication between two physicians. Here we abstract it as an undirected network. The Email network is abstracted from the email communication relationship of the University Rovira I Virgili in Tarragona in the south of Catalonia in Spain. Nodes are users and each edge represents that at least one email was sent.

Table 2. The basic information of networks

Network	*n*	*m*	*dc*	*cc*
Karate	34	78	-0.48	0.57
Risk Map	42	83	0.20	0.52
Dolphin	62	159	-0.04	0.26
Football	115	613	0.16	0.40
Physicians	241	1098	-0.16	0.31
Email	1133	5451	0.08	0.22

4.2 Evaluation metrics

The metrics of Normalized Mutual Information (NMI) [26] and Modularity (Q) [27] are used for evaluating the performance of CSCDA.

A. NMI

NMI is used for evaluating the consistency between communities detected by the algorithm and real communities. A larger NMI represents that communities detected are more consistent with real communities. NMI can be calculated as follows:

$$\mathrm{NMI} = -\frac{2\sum_{i=1}^{C^A}\sum_{j=1}^{C^B} N_{ij}\bullet\log(\frac{N_{ij}\bullet n}{N_i\bullet N_j})}{\sum_{i=1}^{C^A} N_i\bullet\log(\frac{N_i}{n}) + \sum_{j=1}^{C^B} N_j\bullet\log(\frac{N_j}{n})} \tag{8}$$

where A and B are the partitions of real communities and communities detected by the algorithm, C^A is the number of real communities and C^B is the number of communities detected by the algorithm. N_{ij} denotes the number of common nodes in community i of partition A and community j of partition B. N_i and N_j denote the number of nodes of community i and community j, respectively.

B. Modularity

For the networks without real communities, the modularity Q can be used for evaluating the rationality of the communities detected by the algorithm. The modularity Q can be calculated as follows:

$$Q=\frac{1}{2m}\sum_{ij}(A_{ij}-\frac{d(v_i)d(v_j)}{2m})\delta(C_i,C_j) \tag{9}$$

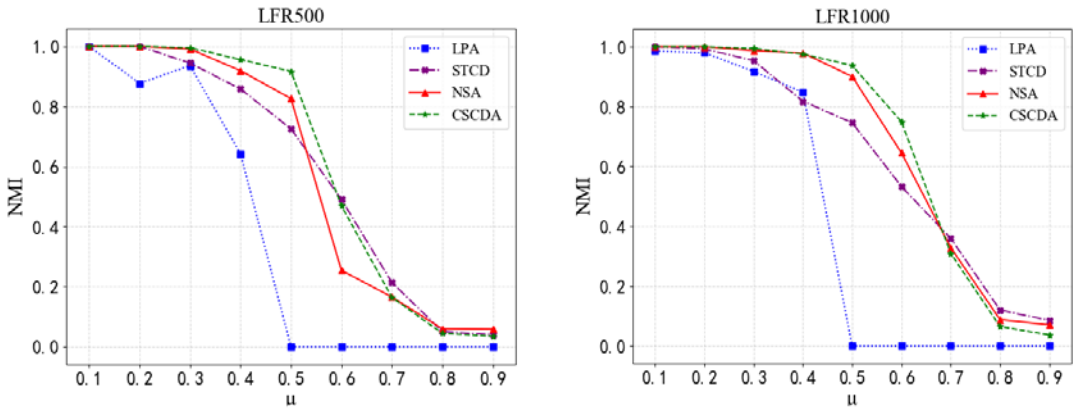
where A_{ij} is the adjacency matrix of the network, C_i and C_j represent the communities of v_i and v_j. If v_i and v_j belong to the same community, then $\delta(C_i,C_j)$=1, otherwise, $\delta(C_i,C_j)$=0.

4.3 Experimental analysis

We compare the result of CSCDA with those of popular algorithms including the label propagation algorithm (LPA) [13], the simplified tree-based community detection algorithm (STCD) [22], and the local node similarity algorithm (NSA) [23]. Due to the randomness of the LPA and STCD algorithms, we run them 10 times on each network. The community metric parameter in NSA is selected through multiple experiments to maximize the modularity of communities of the network.

A. NMI Analysis of LFR Benchmark Networks with Real Communities

The NMI values of four different community detection algorithms on two different scale LFR benchmark networks are shown in **Fig. 2**, where all parameters are fixed except μ, μ is adjusted from 0.1 to 0.9. The larger the value of μ is, the more complex the community structure of the current network is.



(a) The NMI value on LFR500 (b) The NMI value on LFR1000

Fig. 2. The NMI values of different algorithms on LFR benchmark networks

In **Fig. 2(a)**, the NMI values of CSCDA and NSA are 1 for $0.1 \le \mu \le 0.2$, that is, the results of CSCDA and NSA are completely consistent with real communities. Compared with the other three algorithms, CSCDA achieves the optimal NMI result for $0.3 \le \mu \le 0.5$. When $0.6 \le \mu \le 0.7$, the NMI value of CSCDA is only second to that of the best performing STCD. The NMI value of CSCDA is lower than that of NSA and STCD for $\mu \ge 0.8$. However, when $\mu \ge 0.8$, the network structure is extremely complex and does not have obvious communities. At this time, community detection is meaningless. The performance of LPA is not as good as that of the other three algorithms. The NMI value of LPA is all 0 when $\mu \ge 0.5$. The reason is that a huge community is formed by overspreading during the label update process. The performance of CSCDA is better than that of STCD on the whole, and the NMI value is

increased by 2.27% on average. Compared with NSA, the NMI value of CSCDA is increased by 14.21% on average, and the highest increase is up to 84.65% when $\mu = 0.6$.

In **Fig. 2(b)**, CSCDA achieves the best performance in $0.1 \leq \mu \leq 0.5$ as in **Fig. 2(a)** except when $\mu = 0.4$. The difference is that the NMI value of CSCDA is optimal when $\mu = 0.6$, but the improvement is not so obvious. The performance of CSCDA is slightly inferior to that of STCD and NSA when the communities of the network are fuzzy. The NMI value of CSCDA is lower than that of STCD and NSA for $\mu \geq 0.7$. In general, the performance of LPA and STCD is not as good as that of NSA and CSCDA. When $\mu = 0.4$, the NMI value of LPA is higher than that of STCD, which is different from **Fig. 2(a)**. The result may be caused due to the randomness of LPA and STCD. Compared with NSA, the NMI value of CSCDA is increased by an average of 1.91%.

The proposed method, CSCDA, performs best on all LFR benchmark networks during $\mu < 0.6$. As shown in **Fig. 2**, compared with the other three algorithms, CSCDA shows good performance in LFR500 and LFR1000 benchmark networks with real communities on the whole. Especially in small-scale LFR500 benchmark networks, CSCDA has better community detection ability. When $\mu \geq 0.7$, except for the LPA algorithm with the NMI value of 0, the NMI values obtained by other algorithms begin to decrease rapidly. The reason is that when the value of the mixing parameter μ is larger, the topology of the LFR benchmark network becomes more complex and chaotic, thus reducing the quality of the detected communities.

B. NMI Analysis of Real-World Networks with Real Communities

Due to the small sizes of Karate network, Risk Map network and Dolphin network, the detected results can be easily visualized. Next, we will display the detected results and analyze them respectively.

(1) Karate Network

The Karate network reflects the relationship between club members, which was divided into two parts due to disputes between the administrator and the coach. Node ‘1’ and node ‘34’ represent the club's administrator and coach, respectively. The real communities of the Karate network are shown in **Table 3**.

Table 3. The real communities of Karate network

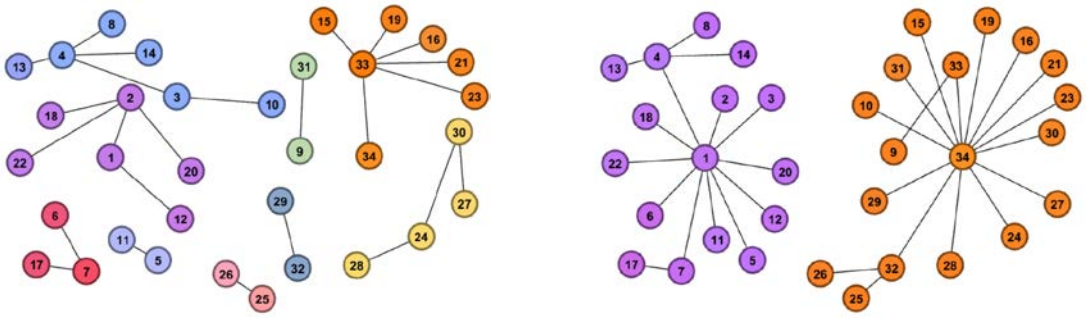
Community ID	Node ID
1	1 2 3 4 5 6 7 8 11 12 13 14 17 18 20 22
2	9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34

According to our proposed method CSCDA, the leadership of each node in the Karate network is first calculated, and then the most closeness similar node is identified for each node in descending order of node leadership. When a node has been identified as the closeness similar node, we no longer consider which node it is the most closeness similar to. Similar nodes are identified for nodes in the Karate network by CSCDA and NSA, and the results are shown in **Table 4**. It can be seen that our proposed method, namely CSCDA, has 72% of the nodes whose most closeness similar nodes are node ‘1’ and node ‘34’. As mentioned above, node ‘1’ and node ‘34’ are two core members of the Karate network. However, only 8% of the nodes are similar to node ‘1’ and node ‘34’ by using the Jaccard index as a similarity measure in NSA. Thus, the proposed method CSCDA can accurately identify the closeness similar node and conform to the actual situation of the real network.

Table 4. The similar nodes of Karate network by CSCDA and NSA

CSCDA				NSA			
Node	The closeness similar node	Node	The closeness similar node	Node	The similar node	Node	The similar node
1	3	10	34	34	33	17	7
34	33	12	1	1	2	18	2
2	1	13	4	3	4	22	2
4	1	14	4	32	29	27	30
6	1	17	7	9	31	15	33
7	1	18	1	14	4	16	33
24	34	20	1	24	30	19	33
32	34	22	1	6	7	21	33
9	33	29	34	8	4	23	33
5	1	27	34	28	24	12	1
11	1	31	34	5	11		
26	32	15	34	20	2		
25	32	16	34	26	25		
30	34	19	34	29	32		
28	34	21	34	10	3		
8	4	23	34	13	4		

Based on the identified similar nodes, each group of similar nodes in the network is merged to construct the initial communities. **Fig. 3** shows the initial communities are formed by the algorithms of NSA and CSCDA on the Karate network. In the forming initial communities, the Jaccard index is used as the similarity measure in NSA. It can be seen that NSA has formed nine initial communities in **Fig. 3(a)**. And **Fig. 3(b)** shows that CSCDA has formed two initial communities by using the closeness similarity measure. We can see that the initial communities detected by CSCDA have been completely consistent with the real communities. The two initial communities are automatically merged, then the modularity of the network is 0. Therefore, the two initial communities are retained as the final communities. The nine initial communities are merged by selecting an appropriate community metric in NSA. Eventually, two communities are detected, however, node '10' is incorrectly divided in NSA. The communities of Karate network detected by CSCDA are shown in **Fig. 4**. The network is naturally divided into two communities, which is completely consistent with real communities, that is, the NMI value is 1.



(a) The initial communities by NSA (b) The initial communities by CSCDA

Fig. 3. The initial communities of Karate network by two algorithms

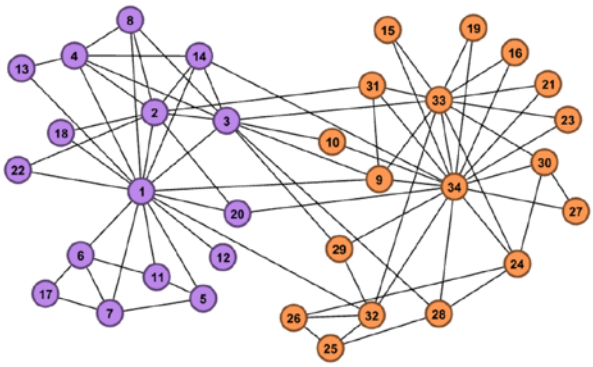


Fig. 4. The communities of Karate network by CSCDA

(2) Risk Map Network

All countries involved in Risk Map network are spread over 6 continents, therefore, the network can be naturally divided into 6 communities. The real communities of the Risk Map network are shown in **Table 5**.

Table 5. The real communities of Risk Map network

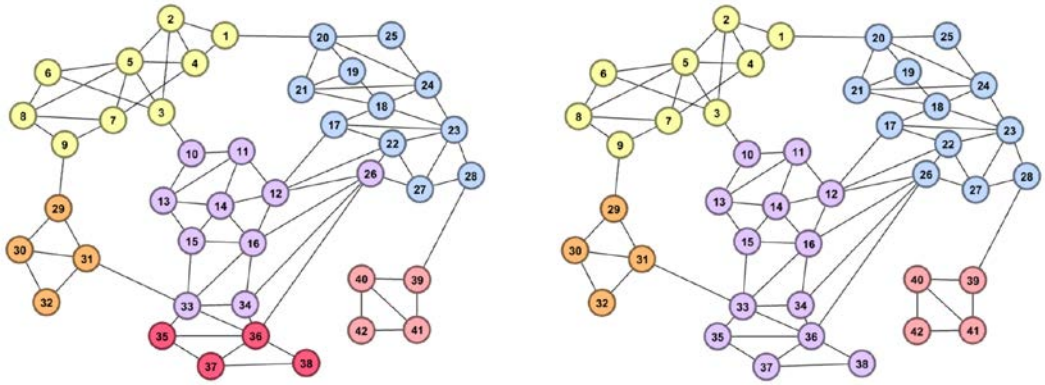
Community ID	Node ID
1	1 2 3 4 5 6 7 8 9
2	10 11 12 13 14 15 16
3	17 18 19 20 21 22 23 24 25 26 27 28
4	29 30 31 32
5	33 34 35 36 37 38
6	39 40 41 42

Fig. 5 shows the community results of the Risk Map network detected by algorithms of NSA and CSCDA. The NSA algorithm divides the network into 6 communities. Although the number of communities is consistent with the real community structure, we can see from **Fig. 5(a)** that node '26' in community ID 3, node '33' and node '34' in community ID 5 are incorrectly divided into community ID 2. From **Fig. 5(b)**, five communities are detected using the proposed algorithm CSCDA. This is because our method adopts the idea of modularity optimization, combining community ID 2 and community ID 5 can get greater modularity. In the case of obtaining high-quality community structure, the communities detected by CSCDA are still closest to the real community structure. Eventually, the NMI value of the community structure that we detected is 0.918.

(3) Dolphin Network

The real communities of the Dolphin network are shown in **Table 6**. The communities detected by CSCDA on the Dolphin network are shown in **Fig. 6**. Except that node '40' is misclassified, the community marked by the purple node is consistent with real community ID 2, and nodes marked by the remaining colors are merged to form a community that is completely consistent with the real community ID 1. At this time, the NMI value is 0.889. CSCDA further divides the remaining color nodes of the Dolphin network into three small communities, so that the communities in the network can obtain the largest modularity. Node '40' has only two first-order neighbor nodes, namely node '37' and node '58'. In the absence of common neighbor nodes, leadership of node '58' is greater than that of node '37', so node '58' is more similar than node '37' to node '40'. Therefore, only qualitative considerations based on topology are taken, without the significance of the actual representation of nodes,

then the communities detected by CSCDA are more reasonable than real communities.



(a) The final communities by NSA (b) The final communities by CSCDA

Fig. 5. The communities of Risk Map network by two algorithms

Table 6. The real communities of Dolphin network

Community ID	Node ID
1	1 3 4 5 9 11 12 13 15 16 17 19 21 22 24 25 29 30 31 34 35 36 37 38 39 40 41 43 44 45 46 47 48 50 51 52 53 54 56 59 60 62
2	2 6 7 8 10 14 18 20 23 26 27 28 32 33 42 49 55 57 58 61

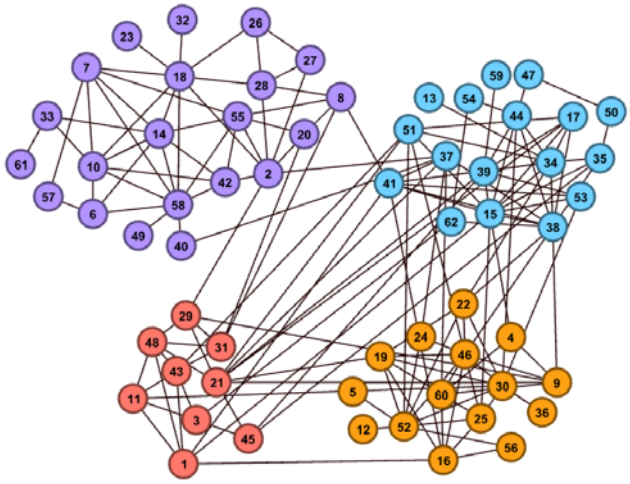


Fig. 6. The communities of Dolphin network by CSCDA

C. NMI and Modularity Analysis of Real-World Networks

Table 7 shows NMI and modularity *Q* comparisons between CSCDA and the other three algorithms on real-world networks. The value of black font is the optimal result and the value of italics is the suboptimal result.

According to the degree correlation of the real-world networks, we divide the networks into two categories, namely disassortative networks (Karate network, Dolphin network, Physicians) and assortative networks (Risk Map network, Football network, Email network).

Table 7. NMI and Modularity comparisons of four algorithms on real-world networks

Network	LPA		STCD		NSA		CSCDA	
	NMI	***Q***	**NMI**	***Q***	**NMI**	***Q***	**NMI**	***Q***
Karate	0.450	0.325	**1.0**	*0.371*	*0.711*	**0.402**	**1.0**	*0.371*
Risk Map	0.821	0.590	0.837	0.589	*0.848*	**0.624**	**0.918**	*0.621*
Dolphin	0.578	0.499	0.616	*0.510*	*0.635*	0.502	**0.889**	**0.523**
Football	0.219	0.552	**0.298**	0.538	0.211	**0.604**	*0.225*	*0.595*
Physicians	/	0.642	/	0.613	/	*0.668*	/	**0.680**
Email	/	0.463	/	0.488	/	**0.540**	/	*0.539*

As shown in the gray part of **Table 7**, CSCDA obtains the optimal NMI value on the Karate network and Dolphin network with real communities, that is, the communities detected by CSCDA are closer to the real communities. And it can be seen that CSCDA achieves the best modularity *Q* on the Dolphin network. Though the modularity *Q* obtained by CSCDA is lower than that of NSA on the Karate network, the result of CSCDA is consistent with the real communities. It is more meaningful to compare the result of CSCDA with the real communities. Compared with the STCD algorithm, the modularity *Q* of our proposed algorithm CSCDA is increased by 10.93% on the Physicians network without real communities. The obtained modularity *Q* still increases by 1.80 %, compared with the NSA algorithm with the suboptimal modularity *Q*.

Risk Map network and Football network are networks with real communities. As shown in **Table 7**, the community structure detected by CSCDA has the highest NMI value in the Risk Map network, and the NMI value is improved by 8.25 % compared with the suboptimal NSA algorithm. Furthermore, CSCDA also achieves suboptimal results in terms of modularity *Q* compared to the other three algorithms. Obviously, the detection results of CSCDA on the Risk Map network are not only the closest to the real communities, but also have a high-quality community structure. And the NMI value and modularity *Q* of CSCDA are suboptimal on the Football network. For the Email network without real communities, the NSA algorithm achieves the optimal modularity *Q*. Although the proposed method CSCDA does not obtain the optimal modularity *Q*, we can see that the modularity *Q* of CSCDA is only 0.2% lower than that of the NSA algorithm.

Therefore, this comparison results show that the proposed method CSCDA outperforms other comparison algorithms on real-world networks as a whole, and can detect reasonable and high-quality communities, especially on the disassortative networks.

Through experiments on the synthetic networks and real-world networks, compared with NSA, the NMI value is increased by an average of 8.06% on the synthetic networks. And in comparison with the optimal values of the other three algorithms, the NMI value is increased by an average of 5.94% on real-world networks with real communities. On real-world networks without real communities, the modularity *Q* is increased by an average of 0.80%. Therefore, especially on the networks with real communities, CSCDA can accurately detect the potential communities.

5. Conclusion

When the current similarity-based community detection algorithms generate the community structure, the detection result is unstable and the accuracy needs to be improved due to the insufficient discrimination of some node pairs. In addition, certain parameters need to be set to obtain the optimal communities. Therefore, we define the common neighbor node clustering

entropy to design a new closeness similarity measure for distinguishing node pairs and propose a noval community detection algorithm, which includes two stages. In the first stage, each node is added to the community where its most closeness similar node belongs through the closeness similarity measure and node leadership. If the most closeness similar node of one node does not belong to a certain community, the node is merged with its most closeness similar node to create an initial community. In the second stage, based on the idea of modularity optimization, the initial communities are optimized by the merging of two levels. The temporary communities with the largest modularity are found through the first-level merging. In the second-level merging, when the temporary modularity no longer increases, the final community detection is completed. The experimental results show that the proposed algorithm CSCDA is superior to the other three algorithms. On the synthetic networks, the community structure detected by CSCDA is closer to the actual community structure, and the real communities can be detected in real-world networks, and at the same time, a higher modularity value can be obtained.

At present, community detection has been deeply researched for traditional static networks, but it still needs further exploration in the diversity and dynamics of networks. In the future, we will explore how to use the closeness similarity measure to analyze the similarity change relationship between nodes and their first-order neighbors, incremental nodes and communities in dynamic networks. Based on this, how to allocate the community for incremental nodes through the idea of modularity optimization and detect a series of dynamic communities will also become the focus of research.

References

[1] X. Li, S. Zhou, J. Liu, G. Lian, and C. W. Lin, "Communities detection in social network based on local edge centrality," *Physica A: Statistical Mechanics and its Applications*, vol. 531, Oct. 2019, Art. no. 121552. Article (CrossRef Link)

[2] W. Liu and L. Chen, "Community detection in disease-gene network based on principal component analysis," *Tsinghua Science & Technology*, vol. 18, no. 5, pp. 454-461, Oct. 2013. Article (CrossRef Link)

[3] P. Chong and B. Shuai, "Measure of hazardous materials transportation network invulnerability based on complex network," *Journal of Central South University*, vol. 45, no. 5, pp. 1715-1723, 2014.

[4] X. Zhou, J. Feng, and Y. Li, "Non-intrusive load decomposition based on CNN-LSTM hybrid deep learning model," *Energy Reports*, vol. 7, pp. 5762-5771, Nov. 2021. Article (CrossRef Link)

[5] M. Liu, J. Guo, and J. Chen, "Community detection in weighted networks based on the similarity of common neighbors," *Journal of Information Processing Systems*, vol. 15, no. 5, pp. 1055-1067, 2019. Article (CrossRef Link)

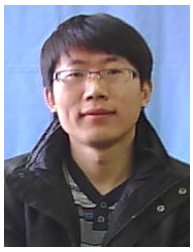
[6] E. Becker, B. Robisson, C. E. Chapple, A. Guénoche, and C. Brun, "Multifunctional proteins revealed by overlapping clustering in protein interaction network," *Bioinformatics*, vol. 58, no. 1, pp. 84-90, Jan. 2012. Article (CrossRef Link)

[7] S. C. Ding, N. Wang, and C. Y. Wu Jing, "Hot topic detection of weibo based on keyword co-occurrence and community discovery," *Journal of Modern Information*, vol. 38, no. 3, pp. 10-18, 2018.

[8] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of web communities," *IEEE Computer*, vol. 35, no. 3, pp. 66-70, Mar. 2002. Article (CrossRef Link)

[9] J. H. Chin and K. Ratnavelu, "A semi-synchronous label propagation algorithm with constraints for community detection in complex networks," *Scientific Reports*, vol. 7, no. 1, pp. 1-12, Apr. 2017. Article (CrossRef Link)

[10] T. Wang, L. Y. Yin, and X. Wang, "A community detection method based on local similarity and degree clustering information," *Phys. A, Stat. Mech. Appl*, vol. 490, pp. 1344-1354, Jan. 2018. Article (CrossRef Link)
[11] H. Tiomokoali and R. Couillet, "Performance analysis of spectral community detection in realistic graph models," in *Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9-18, Mar. 2016. Article (CrossRef Link)
[12] H. Razieh and R. Alireza, "AntLP: ant-based label propagation algorithm for community detection in social networks," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 34-41, Mar. 2020. Article (CrossRef Link)
[13] U. N. Raghavan, A. Réka, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E*, vol. 76, no. 3, pp. 36106-36117, Sep. 2007. Article (CrossRef Link)
[14] Y. Zhang, Y. Liu, J. Zhu, C. Yang, W. Yang, and S. Zhai, "NALPA: A Node Ability Based Label Propagation Algorithm for Community Detection," *IEEE Access*, vol. 8, pp. 46642-46664, Mar. 2020. Article (CrossRef Link)
[15] M. Tang, Q. Pan, Y. Qian, Y. Tian, and X. Wang, "Parallel label propagation algorithm based on weight and random walk," *Mathematical Biosciences and Engineering*, vol. 18, no. 2, pp. 1609-1628, Feb. 2021. Article (CrossRef Link)
[16] C. Wu, Q. Peng, L. Jia, K. Leibnitz, and Y. Xia, "Effective hierarchical clustering based on structural similarities in nearest neighbor graphs," *Knowledge-Based Systems*, vol. 228, no. 4, Sep. 2021, Art. no. 107295. Article (CrossRef Link)
[17] F. D. Zarandi and M. K. Rafsanjani, "Community detection in complex networks using structural similarity," *Physica A: Statistical Mechanics and its Applications*, vol. 503, pp. 882-891, Aug. 2018. Article (CrossRef Link)
[18] Y. Z. Li, H. Xia, R. Zhang, H. B. Xu, and X. G. Cheng, "A Novel Community Detection Algorithm based on Paring, Splitting and Aggregating in Internet of Things," *IEEE Access*, vol. 8, pp. 123938-123951, Jun. 2020. Article (CrossRef Link)
[19] H. Zhang, Y. K. Wu, and Z. Z. Yang, "Community detection method based on multi-layer node similarity," *Computer Science*, vol. 45, no. 1, pp. 216-222, 2018. Article (CrossRef Link)
[20] J. L. Liu, D. L. Wang, S. Feng, and Y. F. Zhang, "Local community detection approach based on fuzzy similarity relation," *Ruan Jian Xue Bao/Journal of Software*, vol. 31, no. 11, pp. 3481−3491, 2020. Article (CrossRef Link)
[21] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and L. Yang, "A local cores-based hierarchical clustering algorithm for data sets with complex structures," *Neural Computing and Applications*, vol. 31, no. 5, pp. 8051–8068, Nov. 2019. Article (CrossRef Link)
[22] L. Bai, J. Liang, H. Du, and Y. Guo, "A novel community detection algorithm based on simplification of complex networks," *Knowledge-Based Systems*, vol. 143, pp. 58-64, Mar. 2018. Article (CrossRef Link)
[23] J. Cheng, X. Su, H. Yang, L. Li, J. Zhang, S. Zhao, and X. Chen, "Neighbor similarity based agglomerative method for community detection in networks," *Complexity*, vol. 2019, pp. 1-16, May. 2019. Article (CrossRef Link)
[24] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E*, vol. 78, no. 4, Oct. 2008, Art. no. 046110. Article (CrossRef Link)
[25] KONECT, Network dataset, 2015, [Online]. Available: http://konect.cc/.
[26] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 9, Sep. 2005, Art. no. 09008. Article (CrossRef Link)
[27] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, no. 6, pp. 66133-66138, Jun. 2004. Article (CrossRef Link)



Wanchang Jiang received a B.S. degree from Liaocheng University, in 2005, and M.S. and Ph.D. degrees from Yanshan University, in 2008 and 2017, respectively. Since 2008, he has been a Teacher with the School of Computer Science, Northeast Electric Power University. Currently, he is an Associate Professor. His research interests include data mining and complex networks.

Xiaoxi Zhang received a B.S. degree from Northeast Electric Power University in 2020. She is currently pursuing a master's degree in the School of Computer Science, Northeast Electric Power University. Her main research interest is community detection.

Weihua Zhu received a M.S. degree from the major of Communication and Information System of Harbin Institute of Technology, in 2007. Currently, he has been a professor with Jilin Technology College of Electronic Information. His research interest is communication and information system.