IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 7, 2019, accepted December 2, 2019, date of publication December 16, 2019,
date of current version January 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2960132

The Multi-Dimensional Information Fusion Community Discovery Based on Topological Potential

RONG FEI[1], SHASHA LI[1], QINGZHENG XU[2], BO HU[3], AND YU TANG[1]
[1]Xi'an University of Technology, Xi'an 710048, China
[2]College of Information and Communication, National University of Defense Technology, Changsha, China
[3]Beijing Huadian Youkong Technology Company, Ltd., Beijing, China

Corresponding author: Shasha Li (xautfei@hotmail.com)

This work was supported in part by the CERNET Innovation Project under Grant NGII20161201, in part by the Scientific and Technological Planning Project of Beilin District of Xi'an under Grant GX1819, in part by the National Key Research and Development Program of China under Grant 2018YFB1201500, in part by the Natural Science Foundation of China under Grant 61773313, and in part by the Teaching Research Foundation of Xi'an University of Technology under Grant XJY1866.

ABSTRACT Many community discovery algorithms add attribute information of nodes to further improve the quality of community division in the complex network with redundant and discrete data, but these algorithms lack of multi-dimensional information, such as users' interests in social networks, social relations, geography and education background, in addition to topological structure and attribute information. Therefore,this paper proposes a Multi-dimensional Information Fusion Community Discovery(MIFCD) method. Firstly, based on the idea of label propagation, link information and attribute information are combined to get link weights between nodes. Secondly, link weights are added to the topology potential to divide the sub group communities. Finally, the sub group communities are combined by using the distance information and attribute information of the core nodes between communities. In order to verify the effectiveness of the algorithm proposed in this paper, the algorithm is compared with six community partition algorithms which only consider the link information of nodes and consider the two kinds of information of node attributes and links. Experiment results on eight social networks show that this method can effectively improve the quality of community classification in both attribute communities and non-attribute communities by analyzing four evaluation indexes: improved modular degree, information entropy, community overlap degree and comprehensive index.

INDEX TERMS Community division, complex network, discrete data, multi-dimensional information.

I. INTRODUCTION

Many complex systems can be regarded as complex and abstract networks composed of vertices and links or edges, such as computer networks, information networks, social networks, biological networks, etc. [1]–[4] Therefore, community detection problems are of great significance to the study of complicated systems and our daily life. In a general manner, community detection views the most closely connected nodes as being part of the same community, so as to better understand the whole social network and utilize resources [5]. In fact, the research results of community detection can be applied to personalized interest recommendation [6], [7], protein function prediction [8], [9], and traffic network detection [10]–[12].

At present, a variety of community detection algorithms have been proposed explore the community structure of complex networks. Based on graph partitioning, community detection needs to define the number of partitions of the community and the volume of the community in advance, realizing the community partitioning by minimizing the number of link edges of a community, such as the Kemighan-Lin algorithm [13] and Spectral Clustering [14]. Community detection based on clustering uses the thought of clustering via the relation of nodes, such as the GN algorithm [15], Newman greedy algorithm [16]. Community detection based on maximum modularity uses the modularity to obtain the optimal network community division, such as the

The associate editor coordinating the review of this manuscript and approving it for publication was Malik Jahan Khan.

Louvain algorithm [17]. The results of these studies is the communities which may incorporate different dimension information since they only consider the strengths of connections between individuals, and there is no analysis of different dimension information characteristics in the process of community detection. However, the network in the real scene, such as social network, will be affected by many factors such as interest, social relationship, region, education background and so on. Therefore, many scholars add the attribute information of nodes into the algorithm research of community detection. Community detection based on a nonnegative matrix uses the thought of a nonnegative matrix, decomposes a node's connection matrix and obtains the node's ownership matrix, such as the SACluster algorithm [18], BAGC algorithm [19] and LANMF algorithm [20]. Community detection based on labels uses randomly generated labels of each node and refreshes the labels of each node in rounds until the labels of each node no longer change, such as the NGLPA algorithm [21], ELPA-ACO algorithm [22], LPPB algorithm [23], etc. Although these algorithms take into account the attribute information of nodes to make the community modular, but due to the characteristics of the real network data, such as redundant relationship, large amount of data storage, discrete data distribution and so on, the community is divided into a high degree of overlap and a large number of communities.Therefore, how to make full use of this complex multi-dimensional information to improve the quality of results has become an important problem for community detection.

The algorithm proposed in this paper divides social networks into communities based on topological potential structure. Community detection based on the domain topological potential uses node connection information to build the topological potential field in which we can partition the community. Many researchers have proposed numerous improved algorithms. For example, the DOCET algorithm [24] is analysed under the topological potential field in the valley structure according to the node position. However, through the experimental process, it is proved that, for the DOCET algorithm, although the value of modularity is large, the number of community partitions is also large.Partitioning the community according to the theory of topological potential causes three or four nodes to be isolated as a community.There are a large number of isolated communities that are easy to affect the public opinion push and community expansion of the real scene.HCDTP algorithm [25] divides the initial community according to the node topology potential, and selects the community corresponding to the maximum module degree as the final community structure by community merging. Although the algorithm reduces the number of isolated communities, it lacks consideration of the network affected by multi-dimensional information because the community members interact in a large number of distinguishable ways in various fields.

Therefore, this paper proposes a Multi-dimensional Information Fusion Community Discovery(MIFCD) method, which is a topological potential community discovery algorithm combined with label propagation. First, the attribute information is constructed in conjunction with the thought of label propagation to obtain the link weight between the nodes. Second, the link weight is added to the topological potential to construct the topological potential field. Third, the core node is used to partition the subgroup community. Fourth, the distance of the core nodes between the subgroup communities is used to partition the community.

The rest of this paper is organized as follows: The second section explains the proper nouns and algorithms appearing. The third section introduces the experimental process and environmental analysis. In the fourth section, the designed algorithm performs confirmatory experiments through multiple experimental sets, including several parameter experiments to optimize the algorithm. The results of the experiment were evaluated by citing several evaluation criteria. Finally, we consider the idea of future development based on these results.The contributions are summarized as follows.

We construct a new community discovery algorithm based on topological potential energy, which is fast and effective for community partitioning with a large number of discrete points.

- We construct attribute characteristics between nodes with link and attribute information of nodes, where communities with closely linked members and highly similar attributes can be detected effectively.
- We design a propagation probability of label, that is the link weight between nodes. It is constructed according to attributes between nodes, which is based on characteristics of Label Propagation Algorithms.
- Community merging is achieved by distance between core nodes and attribute characteristics. It solves the problem of partitioning communities by using the node with the highest local topological potential as the core node of the community. The phenomenon of too many communities and too few nodes in community detection is reduced meanwhile quality of community is kept.

II. RELATED WORKS

A. TOPOLOGICAL POTENTIAL

Topological potential is a virtual potential field constructed in the network topological space, as presented by Gan *et al.* [26]. The topological potential refers to the topology theory in mathematics and the field theory in physics and regards the network G as an abstract system containing n nodes and their interactions. There is a field of action around each nodule, and any node in the field receives the influence of its surrounding nodes. However, the influence of the node rapidly decays as the network distance increases.

Definition 1 (Topological Potential Field): Given a network $G = (V, E)$, all nodes in the network v_i, $1 \leq i \leq n$ have a topological potential $\varphi(v_i)$, and the topological potentials of all nodes interact to form a topological potential field.

Definition 2 (Topological Potential): Given a network $G = (V, E)$, there is a network node $V = \{v_1, v_2, \cdots, v_n\}$ and a node edge set $E = \{(v_i, v_j) \mid v_i, v_j \in V,\ i \neq j\}$, and the topological potential calculation formula of each node is as follows:

$$\varphi(v_i) = \sum_{j=1}^{n} [m(v_j) \times e^{-(\frac{d_{ij}}{\sigma})^2}] \tag{1}$$

where d_{ij} denotes the network distance or hop count between node v_i and node v_j, influencing factors σ are used to control the influence scope of each node, and $m(v_i)$ denotes the weight of node v_i, which is used to describe the inherent attribute of each node. Through similar studies [27]–[30], let $m(v_i) = 1$ in this paper.

According to the mathematical properties of the Gaussian function, if $d_{ij} > \lfloor 3\sigma/\sqrt{2} \rfloor$, the topological potential influence of node v_i to node v_j will rapidly decay to 0 with distance, which can be neglected. The topological potential field is a short-range field whose range of influence is limited. Let $\sigma = 0.4721$ [31] in this paper; then, $\lfloor 3\sigma/\sqrt{2} \rfloor = \lfloor 3 \times 0.4721/\sqrt{2} \rfloor = 1$ which means that network nodes only have influence on their neighbour nodes.

B. LABEL PROPAGATION

Definition 3 (Node Influence): Let each node in the network $G = (V, E)$ have an influence value, expressed in *LR*. Since most of the networks are not connected graphs, this paper uses the LeaderRank algorithm proposed by Lü *et al.* [32] Li *et al.* [33]. to calculate the *LR* values of the nodes. The LeaderRank algorithm mentions that the social network is not a strongly connected graph, so a node g (Ground Node) is introduced to connect with other nodes to make the social network become a strongly connected graph. The core formula of the LeaderRank algorithm is as follows:

$$LR_i(t+1) = \sum_{j=1}^{N+1} \frac{a_{ji}}{K_j^{out}} LR_j(t) \tag{2}$$

$$LR_i = LR_i(t_c) + \frac{LR_g(t_c)}{N} \tag{3}$$

where a_{ji} denotes whether node j to node i has a link (if the link exist, $a_{ji} = 1$, else $a_{ji} = 0$), K_j^{out} denotes the out-degree of node j, N denotes the total number of nodes, $LR_j(t)$ denotes the score of node j at time t, t_c denotes the time of $LR_i(t)$ convergence, $LR_i(t_c)$ denotes the score of the node at time t_c, and LR_i denotes the final score of node i.

Figure 1 shows a small social network topological structure diagram, with 18 nodes in total, which represents a person who has one hobby. We divide the hobbies into two categories and use two different icons to represent people's hobbies. The line connections between nodes represent the relationships between people. As shown in Table 1, through the formula above, we can calculate the node influence of each node of this simple social network dataset.

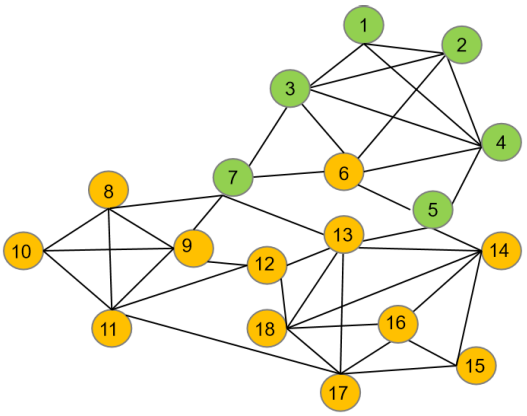


FIGURE 1. Small social network.

TABLE 1. Node influence of small community networks *LR*.

Node ID	Inf	Node ID	Inf
1	0.762913	10	0.762707
2	0.91551	11	1.06772
3	1.06808	12	0.91516
4	1.06807	13	1.37268
5	0.915294	14	1.06761
6	1.06803	15	0.762559
7	1.06789	16	1.06758
8	0.915294	17	1.06758
9	1.06778	18	1.06758

Definition 4 (Propagation Characteristic k): Define $k_{i \leftarrow j}$ as the propagation characteristic metric of the label from node j to node i.

$$k_{i \leftarrow j} = \frac{log(1 + LR_j)}{log((1 + LR_i) \times (1 + LR_j))} \tag{4}$$

This propagation characteristic is determined by the node influence of node v_i and node v_j. While LR_i is much larger than Inf_j, $k_{i \leftarrow j} \approx 1$, which shows that the influence of node v_j is larger, and node v_i will be easily affected by node v_j. In contrast, while LR_j is much larger than LR_i, $k_{i \leftarrow j} \approx 0$, which shows that the influence of node v_i is larger, and node v_i will not be easily affected by node v_j.

From the calculations above, the node influence of each node can be obtained. Figure 2 shows the node influence of node 1, node 2, and node 3. According to the formula of definition 4, we can obtain that:

$$k_{2 \leftarrow 1} = \frac{alog(1 + 0.762913)}{log((1 + 0.762913) \times (1 + 0.91551))} \approx 0.465892$$

$$k_{1 \leftarrow 2} = \frac{log(1 + 0.91551)}{log((1 + 0.762913) \times (1 + 0.91551))} \approx 0.534108$$

$$k_{3 \leftarrow 1} = \frac{log(1 + 0.762913)}{log((1 + 0.762913) \times (1 + 1.0680))} \approx 0.438303$$

$$k_{1 \leftarrow 3} = \frac{log(1 + 1.0680)}{log((1 + 0.762913) \times (1 + 1.0680))} \approx 0.561696$$

The influence LR_1 of node 1 is smaller than the influence of node 3 and node 4. By comparison, we find that the propagation characteristic of node 1 to node 2 is lower than that

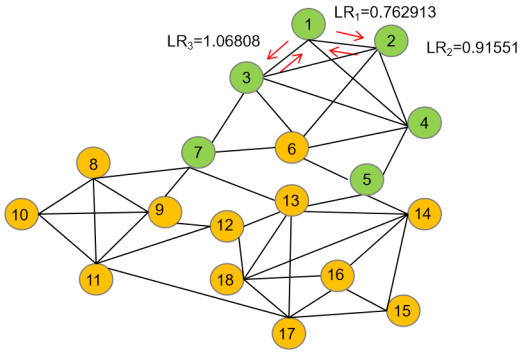


FIGURE 2. Propagation characteristics of nodes in simple social networks.

of node 2 to node 1. Similarly, the propagation characteristic of node 1 to node 3 is also lower than that of node 3 to node 1. Thus, the propagation characteristic value can reflect the difference in the degree of influence between the node with high influence and the node with low influence.

C. SIMILARITY BETWEEN NODES

The realistic social network not only has the topological structure characteristics, but the internal properties of the nodes in the network are also easy to obtain. For example, the scholar records in the C-DBLP have the research direction, the work unit and other information, and thus the attribute characteristics of the nodes (the similarity of the nodes) contain two parts: the structural attribute St and the node internal property In.

$$S_{i,j} = St_{i,j} + In_{i,j} \tag{5}$$

structural attribute:

$$St_{i,j} = \frac{|N_i \bigcap N_j|}{\sqrt{|N_i| \times |N_j|}} \tag{6}$$

node internal property:

$$In_{i,j} = \frac{1}{z}\sum_{k=1}^{z} \zeta(in_{ik}, in_{jk})$$
$$\zeta(in_{ik}, in_{jk}) = \begin{cases} 1, & in_{ik} = in_{jk} \\ 0, & in_{ik} \neq in_{jk} \end{cases} \tag{7}$$

N_i represents the set of all neighbour nodes of node i and node i itself. $in_i = \{in_1, in_2, \cdots, in_z\}$ is the internal property set of node i, in_{iz} is the zth attribute value of the node, and z is the total number of internal attributes.

In the social network data set shown in Figure 1, both node 1 and node 2 have the same neighbour node 3 and node 4, so the structural attribute is $St_{1,2} = \frac{2}{\sqrt{3\times4}} = 0.57735$. Node 1 and node 2 have the same hobby, so the internal property is $In_{1,2} = \frac{1}{2} \times (1 + 1) = 1$. The attribute feature between node 1 and node 2 is $S_{1,2} = 0.57735 + 1 = 1.57735$. In the same manner, $S_{1,3} = 0.51640 + 1 = 1.51640$, $S_{1,4} = 0.51640 + 1 = 1.51640$.

Definition 5 (Transmission Probability of the Label (Correlation Strength Between Nodes, the Edge Weight)): The label of node j propagates to node i with probability $P(i \leftarrow j) \cdot P(i \leftarrow j)$ depends on the similarity measure $S_{i,j}$ of node i and node j, the propagation characteristic metric $k_{i\leftarrow j}$ and the adjacency matrix $\delta(i, j)$

$$P(i \leftarrow j) = S_{i,j} \times k_{i\leftarrow j} \times \delta(i, j) \tag{8}$$

III. THE FRAMEWORK FOR COMMUNITY DISCOVERY

A. IMPROVED TOPOLOGICAL POTENTIAL

In the paper we propose a improved topology potential community discovery algorithm. First, in this work, we use the characteristics of information propagation to transform the attribute structure In and the chain relationship E of the nodes into a link weight relationship R between nodes. Second, we use the topological potential to transform a network structure $G = (V, E)$ with a link relation into a topological potential domain $G' = (V, E, \phi)$ with a mountain shape. Third, we find the local highest point in the spatial structure of the mountain shape, from which we partition the subgroup communities. Finally, according to the distribution of the subgroup community, we merge the subgroup community to obtain the community partitioning result C.

The transmission probability of the label from node i to node j reflects the ability of the label to propagate from node i to node j and can also be considered as the weight of the directed edge of node i to node j. Thus, the weight of the directed edge from node i to node j is formulated as

$$r_{ij} = P(j \leftarrow i) \tag{9}$$

The traditional topological potential algorithm utilizes the link relationship to construct the topological potential field without considering the attribute relationship between the nodules. The definition of a community is to transform a node with a tight link into a community, but the attribute relationship between the nodules also affects the quality of the community partitioning and the application of the real scene.

Therefore, we need to use the attribute relationship and the link relationship between the nodules to construct the environmental impact factor r_{ij} of the topological potential between the nodules; that is, the topological potential between the node i and the node j is affected by the environmental impact factor. The formula for the enhanced topological potential is improved as follows:

$$\varphi(v_i) = \sum_{j=1}^{n}[m(v_j) \times r_{ij} \times e^{-(\frac{d_{ij}}{\sigma})^2}] \tag{10}$$

Since each node has attribute information and link information, it cannot determine the topological potential of the node and judge the impact on the neighbour node simply by the number of links of the node. Thus, we will learn from the thought of label propagation to calculate the probability $P(v_j \leftarrow v_i)$ that the label propagates from node v_j to node v_i and then the environmental impact factors $r_{ij} = P(v_j \leftarrow v_i)$ of nodes v_i and nodes v_j.

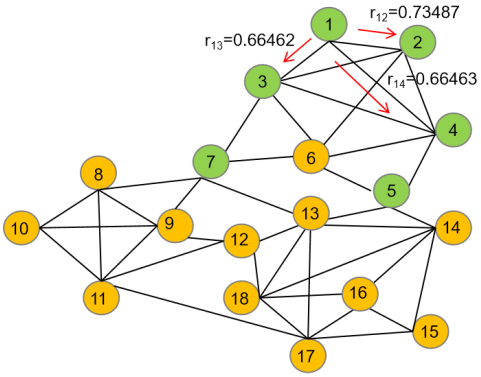


FIGURE 3. Environmental Impact Factor for Node 1 of a Small Social Network.

It can be calculated by the formula above that

$$\begin{aligned} r_{12} &= S_{1,2} \times k_{2\leftarrow 1} \times \delta(1,2) \\ &= 1.57735 \times 0.465892 \times 1 \\ &= 0.73487, \\ r_{13} &= 0.66462, \quad r_{14} = 0.66463. \end{aligned}$$

Due to the topological potential formula of the node

$$\begin{aligned} \varphi(v_i) &= \sum_{j=1}^{n} [m(v_j) \times r_{ij} \times e^{-(\frac{d_{ij}}{\sigma})^2}] \\ &= [\sum_{j=1}^{n} r_{ij}] \times e^{-(\frac{d_{ij}}{\sigma})^2}, \end{aligned}$$

we can calculate $\sum_j^n r_{ij}$ of the node $\sum_j^n r_{ij}$ of v_i first. As shown in Figure 3, what we can know about node 1 is that

$$\begin{aligned} \sum_{j}^{n} r_{1j} &= r_{12} + r_{13} + r_{14} \\ &= 0.73487 + 0.66462 + 0.66463 \\ &= 2.06412. \end{aligned}$$

The transmission probability of the label of node 1 to node 2 determines the capability strength of node 1 to transmit information to node 2, thereby also determining the attribute information and topological potential changed after connection information of node 1 to node 2.

As shown in Table 2, through the formula above, we can calculate result of accumulating the environmental impact factors of each node to its neighbour nodes.

Table 3 is a calculation of the topological potential value of each node of the social network dataset of Figure 1 using the improved topological potential formula 10. The node with the highest local topological potential is marked with a red five-pointed star as shown in Figure 4.

B. SUBGROUP COMMUNITY PARTITIONING

Through the calculation of the node topological potential, the link structure of the network is transformed into the topological potential field of the mountain shape. The partitioning of the community is similar to the partitioning of mountains.

TABLE 2. The sum of the node environmental impact factors for small social networks.

Node ID	$\sum_j^n r_{ij}$	Node ID	$\sum_j^n r_{ij}$
1	2.06412	10	2.06413
2	2.63145	11	3.56152
3	3.43374	12	2.24358
4	3.47960	13	4.06151
5	0.86998	14	3.01314
6	1.05408	15	1.76751
7	0.91799	16	3.94008
8	2.42029	17	3.31787
9	3.27959	18	3.61137

TABLE 3. Node topological potential with link weightfor small social networks.

Node ID	φ	Node ID	φ
1	0.023236	10	0.023236
2	0.029622	11	0.040093
3	0.038654	12	0.025256
4	0.039170	13	0.045721
5	0.009783	14	0.033919
6	0.011866	15	0.019897
7	0.010334	16	0.044354
8	0.027246	17	0.037349
9	0.036918	18	0.040654

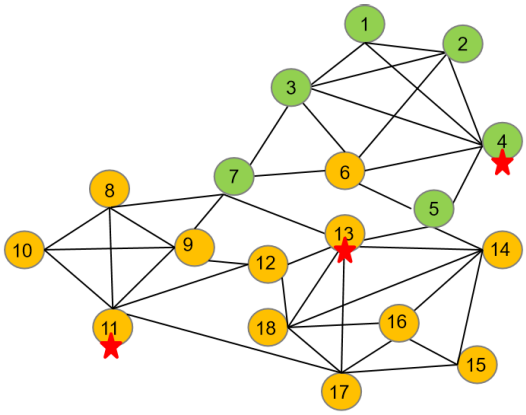


FIGURE 4. Topological potential local maximum node for a simple social network.

Each mountain has peaks, valleys and slopes, corresponding to the core nodes, overlapping nodes and internal nodes of the community. The partitioning of the community is similar to the partitioning of the mountain range, starting from the top of the mountain.

Definition 6 (The Core Nodes): Assume that in a social network $G = (V, E)$, the topological potential field is $G' = (V, E, \phi)$, and the neighbour node of node v_i is N_i. $\forall v_j \in N_i$, and if $\phi(v_i) > \phi(v_j)$, then node v_i is the local highest point of the topological domain.

From the definition above, it can be seen that the core node is the local highest point, that is, the peak node. Because the core node originates from the local maximum value of topological potential, the quality and quantity of community partition through the core node are easily affected. Therefore, the community currently partitioned by the core node is called the subgroup community and needs further processing later. The node identified by the five-pointed star in Figure 4 is the

local highest point of the topological potential, which is the core node of the current subgroup community.

Definition 7 (The Overlapping Node): Assume that, in a social network $G = (V, E)$, the topological potential field is $G' = (V, E, \phi)$, and the neighbour node of node v_i is N_i. $\forall v_j \in N_i$, and if $\phi(v_i) < \phi(v_j)$ and node v_i is in the valley of the community of two different core nodes, then node v_i is the overlapping node of the topological domain, that is, the valley node.

Valley nodes are not necessarily overlapping nodes. If the valley node is located just between the communities of the same core node, the valley node is directly attributable to the community in which its neighbour node is located. Thus, if valley node i is between the communities of two different core nodes, it can be called an overlapping node.

Definition 8 (The Internal Node): Assume that, in a social network $G = (V, E)$, the topological potential field is $G' = (V, E, \phi)$, and the neighbour node of node v_i is N_i. If the internal node satisfies any of the following conditions, then it is tenable: (1) $\exists v_j \in N_i$, if $\phi(v_i) < \phi(v_j)$ and $\exists v_j \in N_i$, and if $\phi(v_i) > \phi(v_j)$, then node v_i is in the slope position, which is the internal node of the topological potential domain. (2) If $\forall v_j \in N_i$, $\phi(v_i) < \phi(v_j)$ and node v_i is in the position of the valley of two communities with the same core node, this node is the internal node of the community.

There are two cases for the internal node. In the first case, the node can be directly judged as an internal node according to its slope position. The second case is more complicated, and we need to judge the nodes of the valley. Because the partitioning of the community is not a parallel process, when the node at the valley location is encountered, only after the neighbour node community of the node is partitioned can we determine whether the node is an internal node or an overlapping node. At the end of the community partitioning, the nodes marked as a valley need to be judged again.

Definition 9 (The Edge Node): Assume that, in a social network $G = (G, E)$, the topological potential field is $G' = (V, E, \phi)$, and the neighbour node of node v_i is N_i. $C_{overlap}$ is a set of overlapping nodes, and $C_{no-overlap}$ is a set of non-overlapping nodes in the community. (1) If $v_i \in C_{overlap}$, then node v_i is an edge node; (2)$\exists v_j \in N_i$, if $v_i \in C_{no-overlap}$ but $N_j \notin C_{no-overlap}$ and $N_j \notin C_{overlap}$, then node v_i is an edge node.

An edge node can be an internal node of a community or an overlapping node of a community. Each node v_i records its shortest distance $CND_{j,i}$ to the core node j of its home community. In theory, most of the edge nodes are farthest away from the core nodes, but it is a bridge of communication between two neighbouring communities. That is, according to the edge node, the core node distance of the two neighbouring communities can be obtained.

The Pseudo code for subgroup community partitioning is given as Algorithm 1. According to algorithm 1, the small social network in Figure 1 is divided into sub group communities and the result is shown in Figure 5.

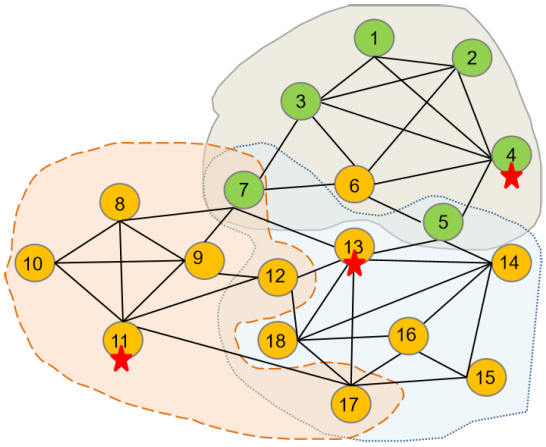


FIGURE 5. Subgroup community partitioning.

Algorithm 1 The Subgroup Community Division

Input: $G' = (V, E, \phi)$, a core node v_k,the neighbour node of node v_k is N_k.

Output: the community node set C_k corresponding to core node k, edge node set C_{k-edge}, internal node $C_{k-internal}$ and overlapping node set $C_{k-overlap}$. $CND_{i,j}$-the distance between the core node i and the node j.

$C_k \leftarrow \phi$, $C_{k-edge} \leftarrow \phi$, $C_{k-internal} \leftarrow \phi$, $C_{k-overlap} \leftarrow \phi$

for all node $v_i \in N_k$ **do**
  if $v_i \notin C_{k-internal}$ and $v_i \notin C_{k-overlap}$ **then**
    $CND_{k,i} \leftarrow 1$
    CommunityExtensionFunction(v_i, $C_{k-internal}$, $C_{k-overlap}$)
  end if
end for
$C_k \leftarrow C_{k-internal} \bigcup C_{k-overlap}$
for all node $v_i \in C_k$ **do**
  if v_i satifies Definition 9 **then**
    $C_{k-edge} \leftarrow C_{k-edge} \bigcup v_i$
  end if
end for
CommunityExtensionFunction(v_i, $C_{k-internal}$, $C_{k-overlap}$)
if v_i satifies Definition 7 **then**
  $C_{k-overlap} \leftarrow C_{k-overlap} \bigcup v_i$
else if v_i satifies Definition 8 **then**
  $C_{k-internal} \leftarrow \bigcup v_i$
  for all node $v_j \in N_i$ **do**
    $CND_{k,j} \leftarrow min(CND_{k,i} + 1, CND_{k,j})$
    CommunityExtensionFunction(v_j, $C_{k-internal}$, $C_{k-overlap}$)
  end for
end if

C. SUBGROUP AMALGAMATION

In the subgroup community partitioning, nodes whose topological potential values are local maximums are considered as peak nodes, and one peak node corresponds to one community. However, there are two situations in subgroup

community partitioning. The first case: If the nodes of the social network dataset are sparse and the degree of nodes is similar, it is easy to cause problems that too many communities will be divided, the community contains too few nodes and so on, which will affect the quality of the community partitioning and the application in reality. Thus, we will make decisions based on the influence of the distance between the peak nodes on the amalgamation of the subgroup. The second case: The divided communities have no path to the others, which means that isolated subgroup communities exist. These isolated subgroup communities cannot be amalgamated by the distance relationship between the core nodes, so we propose a specification to amalgamate isolated subgroup communities to reduce the number of communities.

1) CALCULATE THE DISTANCE BETWEEN SUBGROUP COMMUNITIES

Since the number of nodes in the social network dataset is large, if the distance between the core nodes is calculated by the method of depth-first traversal, then the calculation complexity is high and the time consumption is large; therefore, to obtain the distance between the upper peak nodes quickly, while we partition the sub-group community, we calculate the distance between each node to its community peak node, and finally we analyse three cases to calculate the distance between the subgroup communities.

a: THE TWO SUBGROUP COMMUNITIES DO NOT OVERLAP, BUT THE EDGE NODES ARE CONNECTED(CALCULATE ACCORDING TO EDGE NODES)

As shown in Figure 6, although there are no overlapping nodes in the two subgroup communities, their edge nodes are connected to each other. In this case, since each edge node stores the shortest distance *CND* reaching the subgroup community to which it belongs, we can use the edge nodes to perform information interaction to obtain the distance between the core nodes of the two subgroup communities. However, the distance between the core nodes of the subgroup community to which the edge node belongs is not necessarily the same, and the shortest distance is selected as the distance *CCD* that the two subgroup communities do not overlap but the edge nodes are connected.

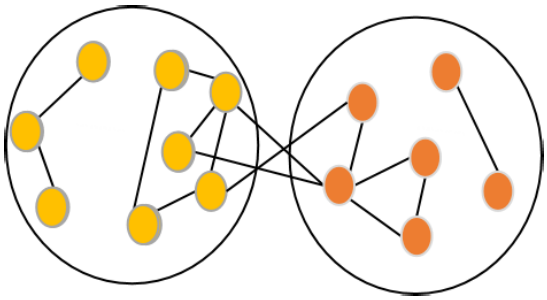


FIGURE 6. Subgroup community is adjacent.

b: SUBGROUP COMMUNITIES DO NOT OVERLAP AND EDGE NODES ARE NOT CONNECTED.(CALCULATE ACCORDING TO EDGE NODES)

The subgroup community is partitioned according to the topological potential value of the node from high to low, but once it hits the current node whose top potential is the local lowest point, that is, when it is partitioned until the valley nodes, the partitioning of current subgroup communities ends. The subgroup community partitioned by this method will actually exhibit the situation shown in Figure 7; that is, the distance between the two sub-group communities is very close, but the topological potential values of the nodes in the middle of their valley nodes are the same, that is, the valley area between the two subgroup communities is a "small plain." The subgroup community partitioning has only reached the nodes on the edge of this "small plain," so the two sub-community communities that do not overlap and the edge nodes that are not connected may have very close distances between their core nodes.

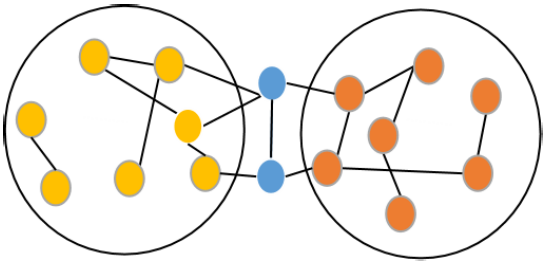


FIGURE 7. Subgroup community is not adjacent.

To calculate the distance between the core nodes of two subgroup communities that do not overlap and edge nodes that are not connected, we use the edge node detection method. The edge node detection method uses the edge nodes of the current subgroup community to jump to the outside of the subgroup community according to the set step size. Whenever the next node is skipped, it is first determined whether the current node belongs to other subgroup communities. If the answer is yes, we calculate the distance between the two communities according to the set step size of the jump and the information of the initial node and the current node; if the answer is no, we jump to the next node. Due to the large number of nodes in the dataset and the complex node relationship, it is impossible to predict whether the closest subgroup community can be found. Therefore, when performing edge detection, the step size of the set is set, and the value of the step size is set to $1/2$ of the Euclidean distance of the current edge node reaching the core node of the subgroup community.

c: SUBGROUP COMMUNITIES OVERLAP (CALCULATE ACCORDING TO OVERLAPPING NODES)

As shown in Figure 8, there are overlapping nodes between subgroup communities, indicating that there is a certain relationship between the two communities. We only need to

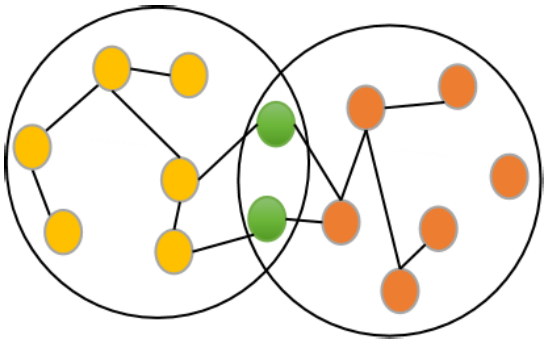


FIGURE 8. Subgroup community overlap.

accumulate the distance from the overlapping nodes of the subgroup community to the core node and finally take the shortest path length.

For the calculation of the distance between subgroup communities in algorithm 2, we first process and calculate the three cases above and obtain the shortest distance of the community. Then, we compare the results of the three cases to take the minimum value. Finally, we obtain the shortest distance between the two neighbouring communities.

2) SUBGROUP COMMUNITY AMALGAMATION

Through the analysis and calculation of the three cases above, we obtain the shortest path of the core nodes between two communities. According to the distance of the core nodes, similar communities can be amalgamated, but in fact, many data sets have a sparse link relationship between nodes, that is, there are many isolated nodes and very small ''isolated'' communities, as shown in Figure 9.

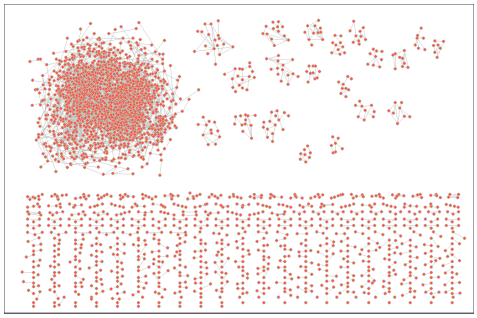


FIGURE 9. Node distribution of the citeseer dataset.

Figure 9 shows the data node distribution of the citeseer data set. From the figure, we can see that the nodes at the top left of the figure have a close relationship, but the nodes below the figure are very sparse. The sparseness of nodes tends to cause the number of partitioned communities to be determined by these sparsely distributed nodes, making the range of community partitioning too small, thus losing the meaning of the community. Therefore, after the subgroup community partitioning, the subgroup community needs to be amalgamated according to the sparse distribution.

Algorithm 2 The Shortest Distance Between Current Core Nodes

Input: the topological potential field $G' = (V, E, \phi)$, the core node set C_{core}, the community node set C_k corresponding to core node k, edge node set C_{k-edge}, internal node $C_{k-internal}$ and overlapping node set $C_{k-overlap}$. CND is the distance between the node and arriving core node, the neighbour node of node v_k is N_k.

Output: the shortest distance between current core nodes CCD.

for all node $v_k \in C_{core}$ **do**
  for all node $v_p \in C_{k-overlap}$ **do**
    if $v_p \in C_{i-overlap}$ and $i \neq k$ **then**
      $CCD_{k,i} \leftarrow CCD_{i,k} \leftarrow min(CND_{k,p} + CND_{i,p}, CCD_{k,i})$
    end if
  end for
end for
for all node $v_k \in C_{core}$ **do**
  for all node $v_e \in C_{k-edge}$ **do**
    for all node $v_n \in N_e$ **do**
      if $v_n \in C_{i-edge}$ and $n \neq k$ **then**
        $CCD_{k,i} \leftarrow CCD_{i,k} \leftarrow min(CND_{k,e} + CND_{i,n} + 1, CCD_{k,i})$
      else if $v_n \notin C_{i-edge}$ and $n \neq k$ **then**
        $d \leftarrow max(CCD)/2$
        $step \leftarrow d - CND_{k,e}$
        NonOverlapCommunitiesDistanceFunction(v_k, v_n, d, $step$)
      end if
    end for
  end for
end for
NonOverlapCommunitiesDistanceFunction(v_k, v_n, d, $step$)
if $step > 0$ **then**
  for all node $v_i \in N_n$ **do**
    if $v_i \in C_{t-edge}$ and $t \neq k$ **then**
      $CCD_{k,t} \leftarrow CCD_{t,k} \leftarrow min(d - step + CND_{t,i} + 1, CCD_{k,t})$
    else
      Function(v_k, v_i, d, $step - 1$)
    end if
  end for
end if

Thus, subgroup community amalgamations are divided into two kinds: (1) adjacent subgroup community amalgamations and (2) sparse subgroup community amalgamations.

a: ADJACENT SUBGROUP COMMUNITY AMALGAMATION

The shortest path of the core node between two adjacent communities is stored in the CCD; calculate $d = max(CCD)$, set φ as the amalgamation parameter to take the value $0 - 1$, and φd is the amalgamation distance. While $CCD_{ij} < \varphi d$,

we perform the community amalgamation, and one core node in the two subgroup communities is randomly set as the core node of the merged community.

b: SPARSE SUBGROUP COMMUNITY AMALGAMATION

Definition 10 (The Sparse Subgroup Community): Assume that, in a social network $G = (V, E)$, C_{core} is the community core node set, C_{k-edge} is the community edge node set of core node k, $maxNum$ is the maximum number of nodes in an isolated community. If $\exists v_j \in C_{core}$, if $\forall v_i \in C_j$, $\forall v_t \in N_i$, $v_t \in C_j$ and $C_j < maxNum$, them the community whose core node j is sparse community.

Since there is no connection between the sparse subgroup community and other communities, it is stipulated that these sparse subgroup communities are amalgamated through the information attributes of their core nodes. That is, the sparse subgroup communities with the same information attributes of the core node points amalgamate into one large community.

Since the nodes of a social network have multiple attributes, there is at least one common attribute between the core nodes of two isolated communities when merging isolated communities.

Algorithm 3 is the pseudo code of sub group community merging, which includes neighborhood community merging and sparse sub group community merging.

Algorithm 3 The Subgroup Community Amalgamation

Input: the topological potential field $G' = (V, E, \phi)$, the core node set C_{core}, the community node set C_k corresponding to core node k, φ as the amalgamation parameter to take the value $0 - 1$, $maxNum$ is the maximum number of nodes in an isolated community, z is the total number of internal attributes .

Output: the Core node set of sparse community C_{sparse}, the subgroup community amalgamation result.

```
/*Adjacent subgroup community amalgamation*/
d = max(CCD)
for all node v_k1 ∈ C_core do
  for all node v_k2 ∈ C_core do
    if CCD_k1,k2 < φd then
      C_core ← C_core − v_k2
      C_k1 ← C_k1 ⋃ C_k2
    end if
  end for
end for
/*Sparse subgroup community amalgamation*/
for all node v_k1 ∈ C_core do
  if C_k < minNum and the community C_k satisfy Definition 10 then
    C_sparse ← C_sparse ⋃ v_k
  end if
end for
for all node v_s1 ∈ C_sparse do
  for all node v_s2 ∈ C_sparse do
    if In_s1,s2 ⩾ 1/z then
      C_core ← C_core − v_s2
      C_s1 ← C_s1 ⋃ C_s2
    end if
  end for
end for
```

IV. EXPERIMENT

A. EXPERIMENTAL DATASET

To prove the improved experiment effective, three community network datasets with links and attributes and five non-attribute social network datasets were selected for the experiment. The specific information is shown in Table 4:

TABLE 4. Dataset information.

Dataset	Nodes Number	Edges Number	Attribute
citeseer	3312	4732	6
cora	2708	5429	5
WebKB	877	1608	5
facebook-combined	4039	88234	-
email-Eu-core	1005	25571	-
p2p-Gnutella06	8717	31525	-
CA-GrQc	5242	28980	-
hep-th	8361	15751	-

B. METHOD OF EVALUATION CRITERIA

To evaluate the improved algorithm, the experiment used the improved modular degree Q^E_{ov} [34], information entropy Entropy [19], community overlap degree Overlap and comprehensive index F as evaluation indexes to observe the quality of community partitioning according to the algorithm.

1) IMPROVED MODULAR DEGREE Q^E_{ov}

Since the main content of this paper is community partitioning of overlapping communities, the evaluation criterion for modular degree is based on the method of introducing an optimization formula of the membership coefficient and simultaneously discovering overlapping and hierarchical community structures. The membership coefficient of the node is redefined as the number of the community to which the node belongs. The higher the improved module value is, the closer the internal links of the community are shown. The specific formula is as follows:

$$Q^E_{ov} = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} [(A_{ij} - \frac{k_i k_j}{2m}) \frac{1}{O_i O_j}] \tag{11}$$

where O_i denotes the number of communities to which node i belongs, and the rest and non-overlapping communities find that the evaluation index module degree Q is similar.

2) INFORMATION ENTROPY

Information entropy uses the formula of the internal nodes of the community to magnify situations of different attributes so as to judge the rationality of the community for

TABLE 5. Parameter φ experiments.

Parameter	Data set	Number of communities	Overlap	Improved modular degree	Information entropy	Comprehensive index F
0.2	citeseer	160	**1.053442**	**0.634315**	0.730994	**0.935933**
	cora	74	1.143279	**0.922984**	**0.624642**	**1.212798**
	WebKB	22	1.080957	0.839338	1.72429	1.292081
0.5	citeseer	132	1.056159	0.612224	**0.714449**	0.909849
	cora	45	**1.138847**	0.563148	0.832312	0.876022
	WebKB	20	**1.079817**	**0.854107**	**1.72246**	**1.309186**
0.8	citeseer	126	1.055253	0.609915	0.707751	0.906295
	cora	43	1.139217	0.559642	0.830639	0.871646
	WebKB	19	1.091220	0.800012	1.64698	1.238248
1.0	citeseer	125	1.065519	0.567909	0.734484	0.863408
	cora	43	1.139217	0.559642	0.830639	0.871646
	WebKB	19	1.091220	0.800012	1.64698	1.238248

attribute partitioning. The larger the information entropy value is, the more situations in which the internal nodes of the partitioned communities have different attributes there are. The analysis of the community partitioning is unreasonable from the perspective of attributes, and thus one hopes that the information entropy value is small. The formula of information entropy is as follows:

$$Entropy = \sum_{i=1}^{z} \sum_{j=1}^{k} \frac{|c_j|}{|V|} entropy(a_i, c_j) \tag{12}$$

where $entropy(a_i, c_j) = -p_{ij} \log_2 p_{ij}$, and p_{ij} denotes the proportion of nodes in community j that have attribute values a_i.

3) COMMUNITY OVERLAP DEGREE *OVERLAP*

The number of overlapping nodes of the community determines the value of the community overlap degree *Overlap*. This value embodies the degree of network coupling and is calculated as follows:

$$Overlap = \frac{1}{m} \sum_{c \in C} |c| \tag{13}$$

where $|c|$ represents the number of nodes of the community c and m represents the number of network nodes.

4) COMPREHENSIVE INDEX *F*

In general, networks with high overlap have relatively low modular degree, and they present negative correlation. For experimental results, the greater the modular degree is, the smaller the information entropy and overlap degree are, and the better the quality of community mining. Therefore, integrating the above situation to output more appropriate community results, the F value is defined as a comprehensive evaluation index.

$$F = \frac{Q_{ov}^{E} \times (Entropy + Overlap) \times 2}{Q_{ov}^{E} + Entropy + Overlap} \tag{14}$$

C. EXPERIMENT

1) PARAMETER EXPERIMENT

In the MIFCD algorithm proposed in this paper, the parameter φ determine the results of the community partitioning of the subgroup community. Therefore, it is necessary to experiment with the parameter φ to determine the value of the parameter. As shown in Table 5, in this experiment, the values of the parameter were set as 0.2, 0.5, 0.8, and 1. Through the experimental results of the parameter of Table 5 and from the perspective of the number of communities, as the parameter increased, the number of communities decreased. However, in the cora dataset and in the WebKB dataset, when the parameter is changed from 0.8 to 1, the number of communities does not change because the complicated distance φd is taken as integer rounding. In terms of the overlap degree, the overlap degree is highest when the parameter in the citeseer dataset is set to 0.2, but the overlap degree is highest when the parameter in the cora dataset and the WebKB dataset is set to 0.5. In terms of improved modular degree, the improved module degree is highest when the parameter in the citeseer dataset and the cora dataset is set to 0.2, but the improved module degree is highest when the parameter in the WebKB dataset is set to 0.5. In terms of information entropy, the improved information entropy is highest when the parameter are set to 0.5 in the citeseer dataset and the WebKB dataset, and the information entropy is highest when the parameter in the cora dataset is set to 0.2. Finally, from the comprehensive index analysis, it is found that the comprehensive index is highest when the parameter of the citeseer dataset and the cora dataset are set to 0.2, but the comprehensive index is highest when the parameter in the WebKB data set is set to 0.5. Therefore, it can be seen that the value of the parameter φ used in this algorithm is most suitable between [0.2, 0.5].

2) EXPERIMENTAL COMPARISON OF ATTRIBUTE DATASETS

In the experiment of attribute datasets, the subgroup community partitioning and amalgamation have been analysed in detail. To better demonstrate the superiority of the algorithm proposed in this paper, the proposed algorithm is compared with the DOCET algorithm, LANMF algorithm, LPPB algorithm, Louvain algorithm, SCD algorithm and DEMON algorithm by way of experiment. The DOCET algorithm, the Louvain algorithm, the SCD algorithm and the DEMON algorithm only consider the link information of

TABLE 6. Subgroup community partitioning of attribute datasets.

Data set	The number of subgroups	The number of isolated subgroups	Overlap	Improved modular degree	Information entropy	Comprehensive index F
citeseer	498	262	1.082125	0.684279	0.743478	0.995442
cora	233	54	1.180206	0.654599	0.885536	0.994164
WebKB	35	4	1.095781	0.819825	1.62086	1.259545

TABLE 7. Subgroup community amalgamation results.

Algorithm	Data set	The number of communities	Overlap	Improved modular degree	Information entropy	Comprehensive index F
MIFCD	citeseer	160	1.0534420	0.634315	0.730994	0.9359330
	cora	74	1.1432791	**0.922984**	0.624642	**1.2127985**
	WebKB	22	1.0809578	**0.839338**	1.72429	**1.2920815**
DOCET	citeseer	646	1.116243	0.613582	0.708683	0.855256
	cora	242	1.245199	0.492039	0.788351	0.732308
	WebKB	42	1.140250	0.721182	1.76545	1.122611
LPPB	citeseer	10	1.2118	0.125659	1.75542	0.241107
	cora	5	1.11041	0.180929	1.3465	0.337038
	WebKB	5	1.05245	0.14025	0.815319	0.260908
LANMF	citeseer	12	1.84662	0.361819	2.42329	0.667109
	cora	10	1.5096	0.442293	1.42948	0.768879
	WebKB	10	1.7252	0.240839	1.79639	0.450845
Louvain	citeseer	462	1.00000	**0.891144**	0.853853	**1.179787**
	cora	105	1.00000	0.820324	0.587519	1.036625
	WebKB	10	1.00000	0.643728	1.50871	0.991060
SCD	citeseer	2006	1.00000	0.422828	0.231667	0.6295375
	cora	1708	1.00000	0.313575	0.0753262	0.4855571
	WebKB	156	1.00000	0.631668	0.50389	0.8896590
DEMON	citeseer	94	**0.295592**	0.229837	**0.17723**	0.3093164
	cora	125	**0.647341**	0.300628	**0.307422**	0.4572735
	WebKB	20	**0.508552**	0.178078	**0.896538**	0.3160948

the nodes in the social network dataset, while the LANMF algorithm and the LPPB algorithm use the link information and attribute information of the nodes in the social network dataset to perform community partitioning. In these four datasets, the DOCET algorithm, the LANMF algorithm, the LPPB algorithm, the SCD algorithm [35] and the DEMON algorithm [36] can partition the overlapping communities, while the Louvain algorithm is mainly for the partitioning of non-overlapping nodes.

a: SUBGROUP COMMUNITY PARTITIONING

This paper performs the subgroup community partitioning on three attribute datasets. First, the core node is determined based on the local highest point of the node topological potential value. Then, the core node is used to partition the subgroup community. Finally, as shown in Table 6, the subgroup community division results are calculated and summarized.

As shown in Table 6, there are 489 subgroup communities in the citeseer dataset in which there are 262 isolated subgroup communities, that is, half of the communities are isolated subgroup communities. The number of nodes in the isolated subgroup community is less than 10, so the number of nodes in half of the subgroup community of the citeseer dataset is too small. The number of subgroup communities in the cora dataset is 233 in which one quarter of the subgroup communities are isolated subgroup communities. The number of the subgroup community in the WebKB dataset whose number of subgroup communities is the smallest compared to the other two attribute datasets but whose comprehensive index is the highest is 35. According to the number of subgroup communities and comprehensive index in the table, the current subgroup partitioning effect of the WebKB dataset is good, while the number of subgroups in the citeseer datasets and the cora dataset is large. The proportion of isolated subgroup communities in subgroup communities is large, and these datasets need to be further amalgamated to ensure the comprehensive quality of community division.

b: SUBGROUP COMMUNITY AMALGAMATION

In the last experiment, the subgroup community of three attribute datasets has been partitioned, and then the community is amalgamated according to the distance *CCD* between the subgroup communities and the complicated distance φd. The value of φ is 0.2, and the result is shown in Table 7.

As shown in Tables 6 and 7, in the citeseer dataset, 498 subgroup communities are amalgamated into 132 communities, which is 1/4 of the number of subgroup communities before amalgamation; in the cora dataset, 423 subgroup communities are amalgamated into 45 communities, which is 1/5 of the number of subgroup communities before amalgamation; in the WebKB dataset, because of the small amount of data, there are 20 communities after the amalgamation, which is 4/7 of the number of subgroup communities before amalgamation. Therefore, as for the algorithm proposed in this paper,

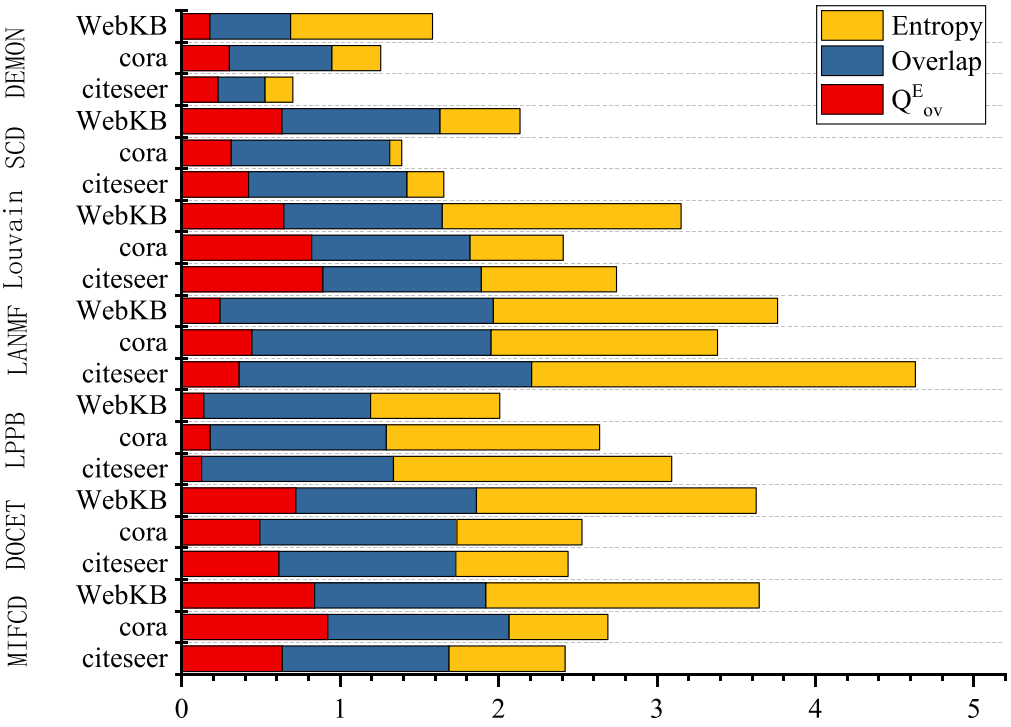


FIGURE 10. Experimental comparison of three evaluation criteria.

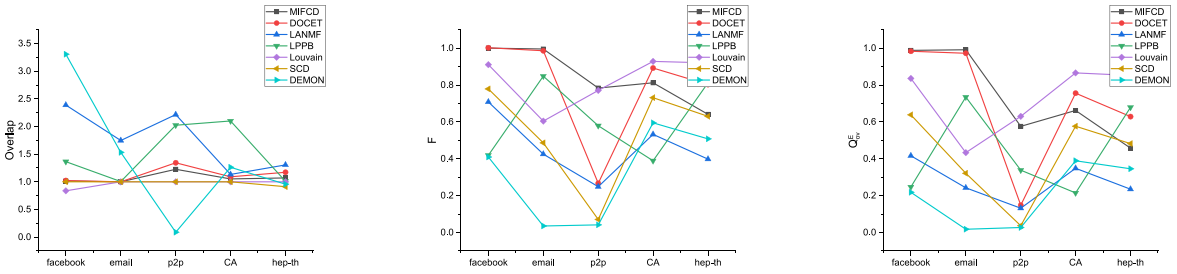
after the community amalgamation, the number of communities in three attribute datasets has decreased. A comparative analysis of the composite index with a parameter value of 0.5 in Table 5 and the composite index in Table 6 found that, in the citeseer dataset, it decreases from 0.959442 to 0.909849 after the amalgamation, while in the cora dataset, it is reduced from 0.994164 to 0.876022 after the amalgamation. The gap between the comprehensive index after and before the amalgamation is approximately 0.1 in the citeseer dataset and the cora dataset. However, the reason for the decrease in the comprehensive index of the two datasets after amalgamation is that the improved module degree of the subgroup communities after amalgamation has decreased. The improved module degree of the citeseer dataset decreases from 0.684279 to 0.612224 after the amalgamation, while the changes in overlap degree and information entropy are not obvious. Likewise, the improved module degree of the cora dataset is also reduced from 0.654599 to 0.563148 after the amalgamation, and the changes in overlap degree and information entropy are not obvious. In contrast, the comprehensive index of the WebKB dataset is higher than the comprehensive index after the amalgamation, rising from 1.255945 to 1.309186, and the difference is approximately 0.05. In the process of the subgroup community amalgamation, the comprehensive index fluctuates approximately 0.1, but the number of communities decreases significantly.

In Table 7, the algorithm proposed in this paper is compared with the algorithm discovered from three other communities. By comparison, it can be seen that, in the citeseer data set, the comprehensive index of the Louvain algorithm is the highest, and the next is the algorithm proposed in this paper. The reason for this is that the Louvain algorithm uses the module degree optimal method to divide the community. Therefore, compared with the improved module degree of the other four algorithms, the Louvain algorithm has the highest improved module degree. Although this paper uses the improved module degree as the evaluation criteria, when the community is a non-overlapping community, the improved module degree formula is actually the module degree formula. Therefore, the improved module degree of the Louvain algorithm is higher than other algorithms, and thus the comprehensive index is also high. However, in the cora dataset and the WebKB dataset, the proposed algorithm is highest in terms of improved module degree and comprehensive index compared with the algorithms found in the other six communities. In this paper, the improved module degree is 0.922984 in the cora dataset, and the improved module degree of the other four algorithms is as low as 0.1; the comprehensive index is 1.2127985, which is at least higher than 0.2 the comprehensive index of the other six algorithms. In this paper, the improved module degree of the WebKB dataset is 0.839338, and the improved module degree of the other six algorithms is at least higher than 0.1; the comprehensive index is 1.2920815, which is also higher than the comprehensive index of the other six algorithms by at least 0.2. Therefore, through the above analysis, the proposed algorithm has certain advantages over the other six algorithms.

Figure 10 shows the improved module degree value, overlap value and information entropy obtained by comparing the

TABLE 8. Number of community data with non-attribute data.

Algorithm / Data set	MIFCD	DOCET	LANMF	LPPB	Louvain	SCD	DEMON
facebook-combined	5	5	5	2	13	695	60
email-Eu-core	2	20	10	5	49	285	3
p2p-Gnutella06	2	112	5	5	62	7919	62
CA-GrQc	33	458	10	5	395	2466	302
hep-th	38	778	5	5	1376	3746	426



(a) Improved modular degree (b) Overlap (c) Comprehensive index

FIGURE 11. Experiment of the non-attribute dataset.

proposed algorithm with the six algorithms in the form of a bar chart. It can be intuitively seen from Fig. 12 that the improved module degree values of the proposed algorithm are higher than 0.5 and that the improved module degrees of the LPPB algorithm, the LANMF algorithm and the DEMON algorithm are all lower than 0.5. In terms of overlap degree, the algorithm proposed in this paper has a similar overlap value to that of the DOCET algorithm, the LPPB algorithm, the Louvain algorithm and the SCD algorithm. The overlap degree of the LANMF algorithm is slightly higher than that of the other six algorithms, and the overlap degree of the DEMON algorithm is the lowest. In terms of information entropy, the information entropy of the proposed algorithm is similar to that of the DOCET algorithm and the Louvain algorithm, while the information entropy of the LANMF is slightly higher than that of other algorithms, and the information entropy of the SCD is the lowest.

3) EXPERIMENTAL COMPARISON OF NON-ATTRIBUTE DATASETS

The algorithm proposed in this paper is based on use of the link information and attribute information to perform the community partition on the social network dataset. However, many studies are based on node-based link information for community partitioning, so we also bring the algorithm into the dataset without attributes for experimentation. Before the experiment, we set the node's attribute type to 1, and the node information is set to the same attribute. In the non-attribute dataset experiment, five datasets were shared and compared with the DOCET algorithm, the LANMF algorithm, the LPPB algorithm and the Louvain algorithm. Table 8 shows the number of community partitions of non-attribute datasets.

Figure 11 shows the result of the improved module degree, overlap degree, and comprehensive index of the non-attribute community partition. According to the improved module degree experiment in Figure 11.a, the proposed algorithm is lower than the DOCET algorithm and the Louvain algorithm in the CA-GrQc dataset and hep-th dataset and only lower than the DOCET algorithm in the p2p-Gnutella06 dataset, while it is higher than the other five algorithms. Finally, the proposed algorithm is higher than the other six algorithms in the facebook-combined dataset and the email-Eu-core dataset. Moreover, as shown in the overlap degree of Figure 11.b, it can be seen that the overlap of the nodes in the DOCET algorithm and the Louvain algorithm is not high, all of which are approximately 1, and the overlap degree value of the DEMON algorithm fluctuates significantly. From the number of communities, the first is in the CA-GrQc dataset, and the number of communities proposed in this paper is 33, which is 1/20 times that of the number of communities of the DOCET algorithm and 1/36 times that of the community of the Louvain algorithm. In the CA-GrQc dataset, the community partitioned by the DOCET algorithm and the Louvain algorithm is approximately 6 nodes on average. Therefore, using the improved module degree to calculate, it will be found that, due to the large number of community partitions, the number of overlapping nodes is small, the improved module degree algorithm is relatively high, and the final evaluation criteria are also the highest. Figure 11.c is a calculation of the comprehensive index, although in Fig, 11.c, the comprehensive index of the proposed algorithm is lower than that of the DOCET algorithm and the Louvain algorithm in the CA-GrQc dataset and the hep-th dataset. The reason for this situation is that the algorithm proposed in this paper has a larger module degree than that of the DOCET algorithm and the Louvain

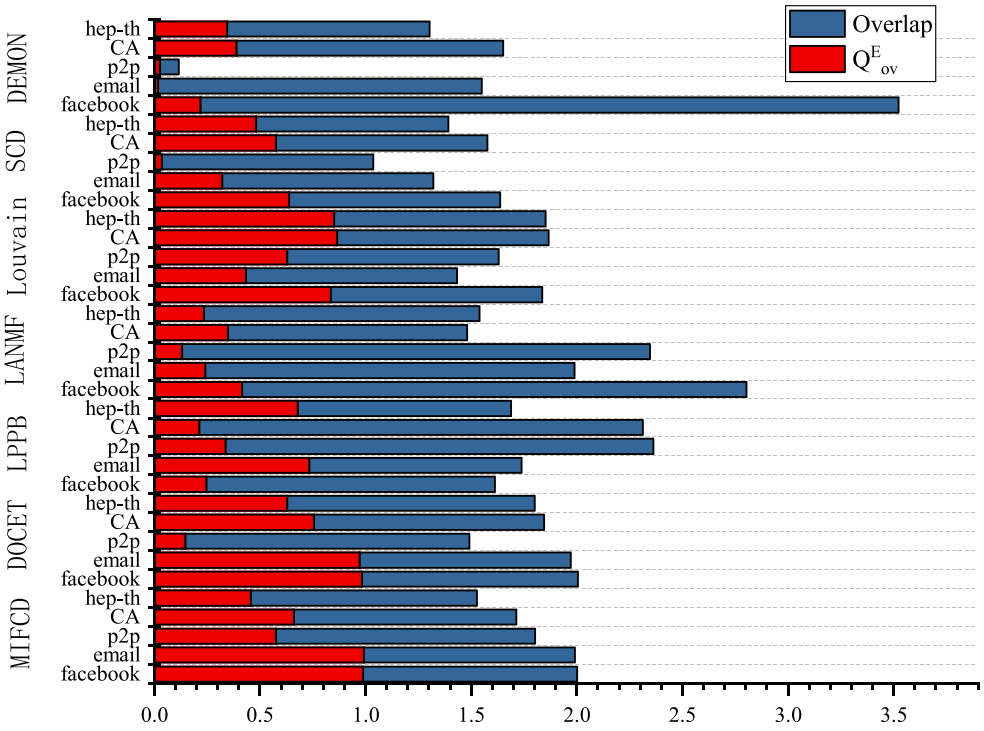


FIGURE 12. Experimental comparison of two evaluation criteria.

algorithm in the CA-GrQc dataset and the hep-th dataset, and the overlap degree is smaller. However, among the remaining three data sets, the proposed algorithm is better than the DOCET algorithm, the LANMF algorithm, the LPPB algorithm, the Louvain algorithm, the SCD algorithm and the DEMON algorithm. Therefore, through the experiment of non-attribute datasets, it is found that the proposed algorithm has certain advantages in the non-attribute datasets. It is also found that the algorithm proposed in this paper is used to experiment on the non-attribute datasets.

Figure 12 is a bar chart of two evaluation criteria for a contrast experiment of seven algorithms in five non-attribute datasets. First, it can be seen from the improved modular degree of the blue bar that the algorithm proposed in this paper makes the improved modular value in the other four datasets higher than 0.5 and an improved modular value lower than 0.5 in the help-th dataset. The other algorithms except the Louvain algorithm with the optimal modular degree whose improved modular degree is higher than 0.5 in the five datasets, and the improved modular degree of the other five algorithms all have the problem with less than 0.5. The improved modular degree of the 5 datasets of the LANMF algorithm is lower than 0.5, and the improved modular degree of the SCD algorithm and the DEMON algorithm in the p2p-Gnutella06 dataset is lower than 0.1. Second, it can be seen from the overlap degree of the red bar that the overlap degree of the algorithm proposed in this paper is similar to that of the DOCET algorithm, the Louvain algorithm and the SCD algorithm, and the overlap degree of the LANMF algorithm and the LPPB algorithm is generally the highest. The overlap degree of the DEMON algorithm is significantly different in the five datasets.

V. CONCLUSION

This paper proposes a Multi-dimensional Information Fusion Community Discovery(MIFCD) method. The algorithm uses the label propagation method to construct the link weights between nodes, so that the nodes in the divided community have close links and the internal attribute characteristics are also highly the same. Because of the characteristics of the actual network data, such as redundant relationship, a large number of data storage, discrete data distribution and so on, the algorithm of community division using local nodes with the highest topological potential as the core nodes of the community is easy to cause high degree of community overlap and large number of communities.Therefore, after the sub-group community is divided, the community merging using the distance between the sub-group nodes and the attribute features solves the above problems while ensuring the tightness of the links between the nodes in the community and the relevance of the attributes.

In order to evaluate the performance of the proposed algorithm, we tested on three attributed data sets and five non-attribute data sets, and and used four evaluation indexes to compare and analyze the DOCET method, the LANMF method LPPB method, the Louvain method, the SCD method and The DEMON method. The experimental results show that the algorithm proposed in this paper has high performance for

the data set with attributes and is also effective for the data set without attributes.

REFERENCES

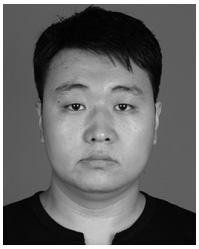
[1] M. Fazil and M. Abulaish, "A hybrid approach for detecting automated spammers in Twitter," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2707–2719, Nov. 2018.
[2] G. Bello-Orgaz, S. Salcedo-Sanz, and D. Camacho, "A multi-objective genetic algorithm for overlapping community detection based on edge encoding," *Inf. Sci.* vol. 462, pp. 290–314, Sep. 2018.
[3] A. Biswas and B. Biswas, "FuzAg: Fuzzy agglomerative community detection by exploring the notion of self-membership," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2568–2577, Oct. 2018.
[4] W. Luo, D. Zhang, H. Jiang, L. Ni, and Y. Hu, "Local community detection with the dynamic membership function," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 3136–3150, Oct. 2018.
[5] J. Sánchez-Oro and A. Duarte, "Iterated Greedy algorithm for performing community detection in social networks," *Future Gener. Comput. Syst.*, vol. 88, pp. 785–791, Nov. 2018.
[6] H. N. Win and K. T. Lynn, "Community detection in Facebook with outlier recognition," in *Proc. 18th IEEE/ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput. (SNPD)*, Jun. 2017, pp. 155–159.
[7] Y.-L. Chen, C.-H. Chuang, and Y.-T. Chiu, "Community detection based on social interactions in a social network," *J. Assoc. Inf.Sci. Technol.*, vol. 65, no. 3, pp. 539–550, 2014.
[8] E. Jaho, M. Karaliopoulos, and I. Stavrakakis, "ISCoDe: A framework for interest similarity-based community detection in social networks," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2011, pp. 912–917.
[9] X. Sun and H. Lin, "Topical community detection from mining user tagging behavior and interest," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 2, pp. 321–333, 2013.
[10] C. Zhong, S. Müller Arisona, X. Huang, M. Batty, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 11, pp. 2178–2199, 2014.
[11] Y. Sun, L. Mburu, and S. Wang, "Analysis of community properties and node properties to understand the structure of the bus transport network," *Phys. A, Stat. Mech. Appl.*, vol. 450, pp. 523–530, May 2016.
[12] L. Sun, X. Ling, K. He, and Q. Tan, "Community structure in traffic zones based on travel demand," *Phys. A, Stat. Mech. Appl.*, vol. 457, pp. 356–363, Sep. 2016.
[13] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Syst. Tech. J.*, vol. 49, no. 2, pp. 291–307, Feb. 1970.
[14] N. F. Michelena and P. Y. Papalambros, "A hypergraph framework for optimal model-based decomposition of design problems," *Comput. Optim. Appl.*, vol. 8, no. 2, pp. 173–196, 1997.
[15] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. Amer. USA*, vol. 99, no. 12, pp. 7821–7826, Apr. 2002.
[16] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, no. 6, p. 066133, 2004.
[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, Oct. 2008, Art. no. P10008.
[18] H. Cheng, Y. Zhou, and J. X. Yu, "Clustering large attributed graphs: A balance between structural and attribute similarities," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 2, p. 12, 2011.
[19] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "A model-based approach to attributed graph clustering," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 505–516.
[20] C. B. He, Y. Tang, H. Liu, G. S. Zhao, Q. M. Chen, and C. Q. Huang, "Method for community mining integrating link and attribute information," *Chin. J. Comput.*, to be published
[21] M. Shen and Z. Ma, "A novel node gravitation-based label propagation algorithm for community detection," *Int. J. Mod. Phys. C*, vol. 30, no. 6, pp. 1–19, 2019.
[22] Y. Li, Y. Zhan, X. Wang, and R. Liu, "Local extended label propagation ant colony optimization overlapping community detection," *Appl. Res. Comput.*, to be published.
[23] S. C. Liu, F. X. Zhu, and L. Gan, "A label-propagation-probability-based algorithm for overlapping community detection," *Chin. J. Comput.*, vol. 39, no. 4, pp. 717–729, 2016.
[24] Z. Wang, Z. Li, G. Yuan, Y. Sun, X. Rui, and X. Xiang, "Tracking the evolution of overlapping communities in dynamic social networks," *Knowl.-Based Syst.*, vol. 157, pp. 81–97, Oct. 2018.
[25] H. Mengnan, W. Zhixiao, H. Jing, R. Xiaobin, and G. Juyuan, "Hierarchical community discovery algorithm for social network on topology potential," *Comput. Eng. Appl.*, vol. 55, pp. 56–63, 2019.
[26] W. Y. Gan, H. E. Nan, L. I. De-Yi, and J. M. Wang, "Community discovery method in networks based on topological potential," *J. Softw.*, vol. 20, no. 8, pp. 2241–2254, 2009.
[27] W. Zhi-Xiao, L. Ze-chao, D. Xiao-fang, and T. Jin-hui, "Overlapping community detection based on node location analysis," *Knowl.-Based Syst.*, vol. 105, pp. 225–235, Aug. 2016.
[28] L. Xiao, "Approach to node ranking in a network based on topology potential," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 33, no. 4, pp. 379–383, 2008.
[29] M. Li, Y. Lu, J. Wang, F.-X. Wu, and Y. Pan, "A topology potential-based method for identifying essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 372–383, Mar./Apr. 2015.
[30] Z. Wang, Y. Zhao, Z. Chen, and Q. Niu, "An improved topology-potential-based community detection algorithm for complex network," *Sci. World J.*, vol. 2014, Jan. 2014, Art. no. 121609.
[31] L. Xiao, S. Wang, and J. Li, "Discovering community membership in biological networks with node topology potential," in *Proc. IEEE Int. Conf. Granular Comput.*, Aug. 2012, pp. 541–546.
[32] L. Lü, Y. C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS ONE*, vol. 6, no. 6, 2011, Art. no. e21202.
[33] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," *Phys. A, Stat. Mech. Appl.*, vol. 404, pp. 47–55, Jun. 2014.
[34] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Phys. A, Stat. Mech. Appl.*, vol. 388, no. 8, pp. 1706–1712, Apr. 2009.
[35] A. Prat-Pérez, D. Dominguez-Sal, and J.-L. Larriba-Pey, "High quality, scalable and parallel community detection for large real graphs," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 225–236.
[36] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: A local-first discovery method for overlapping communities," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 615–623.

RONG FEI received the B.S. and M.S. degrees in computer science and technology and the Ph.D. degree in power electronic and electrical drive from the Xi'an University of Technology, Xi'an, China, in 2002, 2005, and 2009, respectively. From 2009 to 2011, she was a Lecturer with the Xi'an University of Technology. Since 2012, she has been an Associate Professor. Her research interests include community detection, stochastic opposition algorithm, and location-based service.

SHASHA LI received the bachelor's degree from the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China, in 2017. Her research interests include machine learning and network science.

QINGZHENG XU received the B.S. degree in information engineering from the PLA University of Science and Technology, Nanjing, China, in 2002, and the Ph.D. degree in control theory and engineering from the Xi'an University of Technology, Xi'an, China, in 2011.

From 2002 to 2017, he was a Lecturer with the Xi'an Communications Institute, Xi'an. Since 2018, he has been an Associate Professor with the College of Information and Communication, National University of Defense Technology, Xi'an. He is currently a Visiting Scholar with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include opposition-based learning, nature-inspired computation, and combinatorial optimization.



BO HU received the B.S. degree in printing technology and the M.S. degree in signal and information processing from the Xi'an University of Technology, Xi'an, China, in 2005 and 2009, respectively.

From 2009 to 2016, he was an Engineer with the Xi'an Thermal Power Research Institute Company, Ltd., Xi'an. Since 2017, he has been the Director of Research and Development with Beijing Huadian Youkong Technology Company, Ltd. His research interests include image processing and location-based service.

YU TANG is currently pursuing the bachelor's degree in computer science and technology with the Xi'an University of Technology, Xi'an, China. Her research interests include machine learning and intelligent vehicles.