IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 18 December 2016; revised 9 August 2017; accepted 5 September 2017.
Date of publication 11 September 2017; date of current version 9 June 2020.

Digital Object Identifier 10.1109/TETC.2017.2751101

Community Detection by Fuzzy Relations

WENJIAN LUO, (Senior Member, IEEE), ZHENGLONG YAN, CHENYANG BU, AND DAOFU ZHANG

The authors are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, Anhui, China and also with the Anhui Province Key Laboratory of Software Engineering in Computing and Communication, University of Science and Technology of China, Hefei 230027, Anhui, China

CORRESPONDING AUTHOR: W. LUO (wjluo@ustc.edu.cn)

ABSTRACT The increasing demand for knowledge from network data poses significant challenges in many tasks. Discovering community structure from a network is one of the classic and significant problems faced in network analysis. In this paper, we study the network structure from the perspective of the composition of fuzzy relations, and a novel algorithm based on fuzzy relations, i.e., CDFR (Community Detection by Fuzzy Relations), is proposed for non-overlapping community detection. The key idea of CDFR is to find the NGC node (Nearest node with Greater Centrality) for each node and compute the fuzzy relation between them. Then, the community to which a node belongs depends on its NGC node. In addition, the decision graph will be constructed to guide community detection. Experimental results on artificial and real-world networks verify the effectiveness and superiority of our CDFR algorithm.

INDEX TERMS Social network, community detection, fuzzy relation

I. INTRODUCTION

Community detection refers to the recognition of groups from a network, which can be called communities. Thus far, many community detection algorithms have been proposed for this problem, which has received more and more extensive attention. Generally, community detection methods can be categorized into two types: overlapping and non-overlapping.

Overlapping community detection implies that nodes in the network may belong to several communities. Steve Gregory divided such approaches into two types: crisp and fuzzy [1]. The former means that each node belongs fully to each group. The latter indicates that each node of the network does not belong to only one community and is instead affiliated with every community with different membership coefficients. More information regarding overlapping community detection can be found in [2]–[4].

Alternately, non-overlapping community detection is a significant topic in the analysis of network structure. Non-overlapping community detection refers to the case in which the nodes in the network are divided into several communities, and each node belongs to only one community. In other words, there are no intersections among the communities recognized by a non-overlapping community detection algorithm.

In general, community detection can be considered an extension of clustering in social networks. Thus, there are similarities between the approaches for studying these two problems. Community detection algorithms can draw inspiration from clustering algorithms. Recently, a state-of-the-art clustering algorithm named FDP was proposed by Rodriguez *et al.* in [5], which attracted much attention. The main idea of FDP includes the following points: 1) for any data object x, its cluster label depends on the nearest data object (e.g., y) of x that has a larger density, and 2) the central data object of a cluster is far away from its y. Similarly, in a social network, for each node v, its nearest node with larger centrality may have important influence on the community label of node v.

In a social network, the community center node is the core and the most influential node of a community; it is the key node for constructing the community. Intuitively, the centrality of a node represents an influence on the surrounding environment; it can also be regarded as attraction. Therefore, it seems that a node with larger centrality has a greater influence on the surrounding nodes. In this paper, the node with the largest centrality is considered to be the community center. However, the centrality is not the only criterion for identifying the community central node. Although some nodes have large centrality, they cannot be center nodes because they are tightly tied to a node with larger centrality. In other words, their influence is covered by the more powerful nodes. Thus, if a node wants to be a center node, it should be far away from any more powerful node. A larger distance to a more powerful node can be regarded as a smaller dependence on it. Overall, a

node with smaller dependence on more powerful nodes is more likely to be a potential center node of a community.

Alternately, many algorithms have been proposed for non-overlapping community detection, but few algorithms detect communities from the perspective of fuzzy concepts. Most community detection algorithms based on fuzzy concepts are overlapping methods [6]. To the best of our knowledge, only [7], [8] are related to fuzzy concepts and recognize non-overlapping communities, which will be introduced in the next section.

Therefore, in this paper, a novel community detection algorithm named Community Detection by Fuzzy Relations (CDFR) is proposed to discover non-overlapping communities. The core idea of our algorithm is based on the following points:

(1) For each node, its NGC node is defined as the Nearest node with Greater Centrality, which plays an important role in this algorithm.
(2) The dependence of each node on its NGC node is abstracted as a fuzzy relation, which is calculated by use of the composition of fuzzy relations.
(3) Finally, each node can be affiliated with the community that its NGC node belongs to if the fuzzy relation is large enough. In contrast, if the fuzzy relation is small, then that node is considered to be the center node of a new community.

The rest of our paper is organized as follows. An overview of various classical community detection algorithms and typical related algorithms is summarized in Section II. After that, a detailed introduction to our algorithm is provided in Section III. Experiments and comparisons are given in Section IV. Furthermore, discussion regarding CDFR is explored in Section V. Finally, we summarize our work in Section VI.

II. RELATED WORKS

In this section, some classical non-overlapping community detection algorithms are reviewed. Some typical algorithms related to CDFR are also introduced briefly.

The KL algorithm [9] is one of the classical algorithms based on graph partition. In the KL algorithm, the network data can be divided into several groups by cutting off some edges with the purpose of minimizing the total cost of all truncated edges. Other similar methods of this type are discussed in [10], [11].

Girvan and Newman proposed a series of community detection algorithms. In 2004, they proposed the GN algorithm [12], which focuses on edges instead of nodes. The edges with higher edge-betweenness are more likely located between two communities. The communities can finally be recognized by deleting these edges. In addition, many improved algorithms based on GN can be found in [13]–[17].

Newman and Girvan originally proposed the concept of modularity Q [18], which is by far the most popular metric for evaluating the quality of community structure. An algorithm with more modularity is believed to have better performance. Since then, many algorithms have made great efforts to optimize modularity Q. In 2004, Newman proposed [19] to maximize modularity. The core idea is that nodes are aggregated continuously to construct communities if the modularity gain is positive. In addition, FastModularity proposed by Clauset *et al.* in [20] adopts a greedy strategy to optimize modularity. Louvain, proposed by Blondel in [21], is another famous method.

Raghavan *et al.* proposed the Label Propagation Algorithm (i.e., LPA) in 2007 [22]. In the initial situation, each node in the network gets a unique label. Then, the label of each node is refreshed round by round. For each node V, it will obtain the label that appears most frequently in V's neighbors. Since then, many improved versions have been proposed (e.g., SLPA [23], WLPA [24], COPRA [25], and BMPLA [26]).

In addition, some algorithms based on fuzzy concepts have been studied. Sun *et al.* proposed a method based on the composition of fuzzy relations for community detection [8]. It posits that the connection relation between two nodes in a network is considered a fuzzy relation. First, it constructs a fuzzy relation matrix R with the similarities of nodes. Furthermore, the fuzzy equivalence relation $t(R)$ is obtained by use of the composition of fuzzy relations iteratively. Finally, after the fuzzy relation is stable, they map the communities as equivalence classes, which generate the community structure.

After that, an extension of the previous article was proposed by Sun *et al.* [7], and their core ideas are similar. The main difference is that the similarity between nodes is based on edge centrality in [7], while in [8], the intersection of the k-distance neighborhood is regarded as the similarity. However, both [8] and [7] are sensitive to a parameter λ.

You *et al.* proposed IsoFdp in 2015 [27], which extended FDP to detect communities in networks. Calculating the distance between two nodes is the key of this algorithm. Because the distance in network is quite different from spatial data, the network is embed into a low-dimensional manifold, and geodesic distance is used. However, the drawback of this algorithm is that the time complexity is too high.

Recently, some community detection algorithms based on central nodes have been proposed. Khorasgani proposed a novel community detection algorithm named Top-Leader [28] based on the k-means clustering algorithm. The main idea of Top-Leader includes the following points: first, the promising leaders in the social network are recognized; after that, the follower nodes will be assigned to their nearest leader nodes; then, the leader node of each community will be recalculated. The process repeats until convergence.

The LICOD [29] algorithm based on central nodes was proposed by Yakoubi *et al.* At first, many leader nodes are recognized, and if the intersection of two leader nodes is large enough, they should be affiliated with the same group. After that, calculate the membership of each node to each group. Finally, each node will be assigned to the group with the largest membership.

The Leader-Followers [30] algorithm was proposed by Shah. After the leader nodes are selected, the so-called loyal follower nodes are assigned to them.

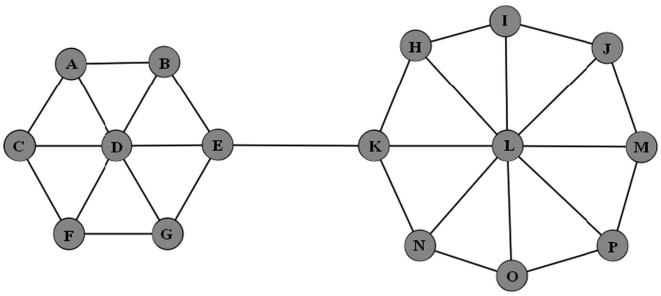


FIGURE 1. An example of a network.

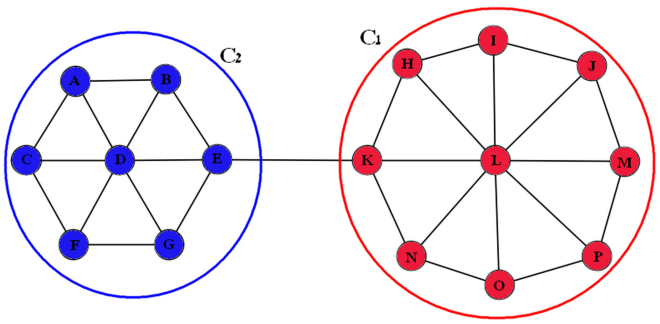


FIGURE 2. The community structure recognized.

In this paper, a novel algorithm (i.e., CDFR) based on fuzzy relations for discovering non-overlapping communities is proposed. However, the fuzzy relation used in this paper is different from the fuzzy relation matrix in the work of Sun. In addition, although both CDFR and IsoFdp are inspired by the clustering algorithm FDP, the complexity of CDFR is lower than IsoFdp, and the ideas of these algorithms are different.

III. THE PROPOSED ALGORITHM

In this section, the proposed algorithm CDFR is introduced in detail. First, the concept of an NGC node is proposed, and the algorithm for finding NGC nodes and computing fuzzy relations is given. Then, the decision graph and its role are introduced. Finally, the complexity of CDFR is estimated.

A. THE NGC NODE AND THE PRIMARY IDEA

The NGC node refers to the Nearest node with Greater Centrality (abbreviated as NGC node), which is a very important type of node. The role of the NGC node is to determine the community label of each node.

The primary idea of our algorithm is given as follows. First, find the NGC node (i.e., NGC(V)) of each node (e.g., V) in the network. Then, define the fuzzy relation between V and NGC(V), which can be regarded as the dependence of V on NGC(V). After that, all the nodes will be sorted in descending order based on their centrality, and the nodes with greater centrality will be processed preferentially. Finally, each node can be affiliated with the community that its NGC node belongs to if the fuzzy relation is large enough. In contrast, if the fuzzy relation is small, then that node is considered a center node of a new community.

Figures 1, 2 and Table I are used to explain the role of NGC nodes and the primary idea. In the network of Figure 1, nodes A-P constitute a network that contains 2 communities (i.e., $C_2 = \{A, B, C, D, E, F, G\}$ and $C_1 = \{H, I, J, K, L, M, N, O, P\}$). For the sake of simplicity, we assume that the degree of a node is considered as its centrality and that the hop count between two nodes is regarded as the distance. A small distance implies a large fuzzy relation; on the contrary, the fuzzy relation is small for a large distance. Table I shows the NGC nodes and the distance from each node to its NGC node.

First, node L with the greatest centrality has no NGC node, which means that it is the first node to be processed and that it is the center node of the community C_1. After that, node D has relatively large centrality and will be processed in the next step. The distance from node D to its NGC node (i.e., L) is 3, which is too large. Thus, node D is considered to be the center node of another community C_2. However, nodes H, I, J, K, M, N, O, and P are close to their NGC node (i.e., the distances to the NGC node L are 1), so they are affiliated with community C_1. Similarly, the distances from nodes A, B, C, E, F, and G to their NGC node are 1, which are small enough. Thus, they obtain the community number (i.e., C_2) from their NGC node (i.e., D). Figure 2 shows the community structure recognized by our primary idea.

B. THE CENTRALITY

However, directly regarding the degree as the centrality of a node is a rough method. Considering that the micro environment composed of two (or more) layer neighbors can reflect more properly the centrality of a node, we use the following formula to calculate the centrality:

$$centrality(V) = deg(V) + \sum_{U \in \Gamma(V)} \left(deg(U) + \sum_{W \in \Gamma(U)} deg(W) \right) \tag{1}$$

where V is any node in the network and $\Gamma(x)$ is the set of direct neighbors of node x. $deg(x)$ is the degree of node x.

The centrality of a node can be regarded as the importance or influence of the node in the network. We care about the value of centrality, as well as the ranking of each node in the network. There are many ways to calculate the degree of centrality. For example, the degree (i.e., the number of direct neighbors) can be taken as the centrality directly. However, many nodes could have the same centrality, and there is no difference between them. Alternately, other methods of calculating the centrality are time-consuming, e.g., the method based on the shortest path or betweenness [31], [32]. Considering the above two factors, the above calculation method of

TABLE 1. The NGC nodes and the distances to their NGC nodes.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Centrality	3	3	3	6	4	3	3	3	3	3	4	8	3	3	3	3
NGC node	D	D	D	L	D	D	D	L	L	L	L	-	L	L	L	L
Distance to the NGC node	1	1	1	**3**	1	1	1	1	1	1	1	-	1	1	1	1

the centrality is used. The advantages of this method are: 1) the centralities of different nodes tends to be distinguishable; and 2) the complexity is relatively low.

C. FUZZY RELATION

For any two nodes V and U, the fuzzy relation $R(V, U)$ is defined as "the followership or dependence relation from node V to node U." Node V belongs to the community of node U if the fuzzy relation is large enough. In contrast, if the fuzzy relation is small, they should be assigned to different communities. This type of fuzzy relation $R(V, U)$ is asymmetrical, which means that node V is close to node U, but this does not mean that the relationship from node V to U is close.

This is consistent with the actual situation in the social network. For example, an employee of a company considers his department leader frequently, which means that R (employee, leader) is relatively large. However, because the leader is responsible for many things, he/she may not have much time to focus on employees. Thus, R(leader, employee) could be small.

Suppose that V is any node of a network and that NGC(V) refers to the NGC node of V. Now, we need to measure the fuzzy relation from V to $\mathrm{NGC}(V)$. A large fuzzy relation from V to $\mathrm{NGC}(V)$ indicates that V should be affiliated with the community that $\mathrm{NGC}(V)$ belongs to. In contrast, V is the beginning of a new community that is different from $\mathrm{NGC}(V)'$s if the fuzzy relation is large.

There are many paths from node V to node $\mathrm{NGC}(V)$. Suppose that P is a set of all paths from V to $\mathrm{NGC}(V)$. p is one path from P, which can be expressed as $p = \{N_1, N_2, \ldots, N_k\}$, where N_i refers to a node in the path p. Obviously, N_1 is V, and N_k is $\mathrm{NGC}(V)$. According to the law of the composition of fuzzy relations [33], [34], the fuzzy relation from V to $\mathrm{NGC}(V)$ should be defined as:

$$R(V, NGC(V)) = \max_{p \in P}\{\mu_p(N_1, N_k)\} \tag{2}$$

where $\mu_p(N_1, N_k)$ can be calculated with the following formula:

$$\mu_p(N_1, N_k) = t(\mu_p(N_1, N_{k-1}),\ \mu_p(N_{k-1}, N_k)) \tag{3}$$

where $\mu(N_i, N_{i+1})$ can be calculated as follows ($\Gamma(x)$ indicates the neighborhood of node x):

$$\mu_p(N_i, N_{i+1}) = \frac{1 + |\Gamma(N_i) \cap \Gamma(N_{i+1})|}{|\Gamma(N_i)|} \tag{4}$$

$t[.]$ is the t-norm, which has two typical forms: Minimum and Product. In this paper, the latter is adopted, i.e.,

$$t(x, y) = x \cdot y \tag{5}$$

Here is an example to illustrate the fuzzy relation between D and NGC(D).

As shown in Figure 3, suppose that we need to find the NGC node of D and calculate $R(D, \mathrm{NGC}(D))$. Obviously, L is the only node with larger centrality than D, and L is the NGC node of D. There are many paths from D to L. For the path $p = \{D, E, K, L\}$, Formula (4) is used to calculate the intersection rate of two adjacent nodes. That is, $\mu_p(D, E) = 3/6$, $\mu_p(E, K) = 1/4$, $\mu_p(K, L) = 3/4$. Therefore, $\mu_p(D, L) = (3/6) * (1/4) * (3/4) = 3/32$. However, there are other paths from D to L, but $p = \{D, E, K, L\}$ obtains the largest $\mu_p(D, L)$. Thus, the fuzzy relation between D and L is 3/32, i.e., $R(D, \mathrm{NGC}(D)) = 3/32$.

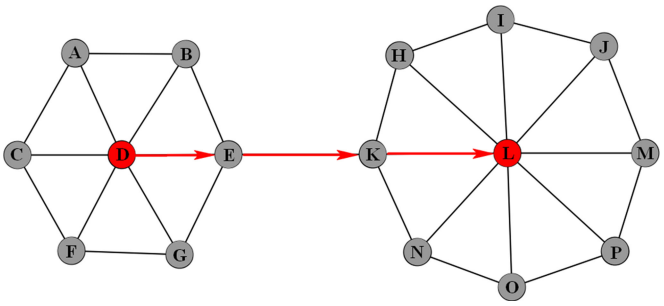


FIGURE 3. **Compute the fuzzy relation** $\mathrm{R}(D, \mathrm{NGC}(D))$.

In Figure 3, D is the NGC node of A. $R(A, D)$ is 1, while $R(D, A)$ is 0.5. This means that node A tends to follow node D completely. However, node D does not concern A so much because it has other connections.

D. FIND THE NGC NODE

Here, an algorithm is proposed to find the NGC node and calculate the fuzzy relation between any node (e.g., V) and its NGC node (e.g., W).

Algorithm 1 shows the main procedure for finding the NGC node of V and calculating the corresponding fuzzy relation. In this algorithm, W is used to represent NGC(V); *fuzzyrelation* is used to store the fuzzy relation $R(V, W)$; for each node X, X.*centrality* denotes the centrality of X, and X.*mju* is used to temporarily preserve the fuzzy relation from V to X. In addition, two data structures, *OpenTable* and *CloseTable*, are required. The former is used to store nodes that are found but not visited, while the latter stores nodes that have been visited.

Algorithm 1 includes the following steps.

(1) Set a flag *findtag* to be false, which is true if the NGC node is found and false otherwise.

(2) Add each direct neighbor (e.g., X) of V into *OpenTable*, and the intersection rate between V and X is regarded as X.*mju*.

(3) Get a node C from *OpenTable*: for any node X in *OpenTable*, C.*mju* is no less than X.*mju*. If *findtag* is false and C.*centrality* is greater than V.*centrality*, node C is the nearest node with the greater centrality found thus far. Thus node C will be the NGC(V) if no other nodes' fuzzy relations are the same as C.*mju*. However, when NGC(V) is not unique, the node with the greatest centrality will be the NGC(V). Once NGC(V) is determined, the algorithm will terminate.

(4) Remove C from *OpenTable* and add it to *CloseTable*. For each direct neighbor of node C (i.e., Y), calculate the one-way intersection rate (i.e., $\mu(C, Y)$) from C to Y with Formula (4). Let *currentfr* take the value $t\{\mu(V, C), \mu(C, Y)\}$ temporally.

Algorithm 1. Finding the NGC Node and Calculating the Corresponding Fuzzy Relation.

Input: Network: G
Node: V
Output: The NGC node of V: W
The fuzzy relation from V to W: *fuzzyrelation*

$W \leftarrow V$
Fuzzyrelation $\leftarrow 0$
OpenTable $\leftarrow \emptyset$
CloseTable $\leftarrow \emptyset$
findtag $\leftarrow$ False
For each X in $\Gamma(V)$ **do**:
  $\mu(V,X) \leftarrow \frac{1+|\Gamma(V)\bigcap\Gamma(X)|}{|\Gamma(V)|}$
  $X.mju \leftarrow \mu(V, X)$
  Add X to *OpenTable*
End For
While *OpenTable* is not empty **do**:
  $C \leftarrow$ the node X with the max $X.mju$ in *OpenTable*
  If *findtag* is False **then**:
    If $C.centrality > V.centrality$ **then**:
      $W \leftarrow C$
      Fuzzyrelation $\leftarrow C.mju$
      Findtag $\leftarrow$ True
    End If
  Else:
    If $C.mju < fuzzyrelation$ **then**:
      break
    End If
    If $C.centrality > W.centrality$ **then**:
      $W \leftarrow C$
      Fuzzyrelation $\leftarrow C.mju$
    End If
  End If
  Add C to *CloseTable*
  Delete C from *OpenTable*
  For Y in $\Gamma(C)$ **do**:
    $\mu(C,Y) \leftarrow \frac{1+|\Gamma(C)\bigcap\Gamma(Y)|}{|\Gamma(C)|}$
    $currentfr \leftarrow C.mju \times \mu(C,Y)$
    If Y not in *OpenTable* **and** *CloseTable* **then**:
      $Y.mju \leftarrow currentfr$
      Add Y into *OpenTable*
    Else If Y in *OpenTable* **then**:
      If $currentfr > Y.mju$ **then**:
        $Y.mju \leftarrow currentfr$
      End If
    Else If Y in *CloseTable* **then**:
      If $currentfr > Y.mju$ **then**:
        $Y.mju \leftarrow currentfr$
        Add Y to *OpenTable*
        Delete Y from *CloseTable*
      End If
    End If
  End For
End While
Return W, *fuzzyrelation*

After that, three situations should be taken into consideration.

First, Y will be added into *OpenTable*, and $Y.mju$ is assigned as *currentfr* if Y is not in *OpenTable* and *CloseTable*.

Second, if Y is in *OpenTable*, update $Y.mju$ with *currentfr* if *currentfr* is larger.

Third, if Y is in *CloseTable* and *currentfr* is larger than $Y.mju$, update $Y.mju$ with *currentfr*, delete Y from *CloseTable* and add Y to *OpenTable*.

(5) Jump to step (3) until *OpenTable* is empty.

It is noteworthy that the node with the largest centrality in network G has no NGC node, so we assume its NGC node is itself, and the fuzzy relation from this node to its NGC node is 0. Obviously, as the node with the largest centrality in the network, it is the center node of the first identified community.

E. FUZZY RELATION WITH REFINEMENT

The node with greater centrality and smaller fuzzy followership relation to its NGC node is more likely to be a community center. Thus, the gap in fuzzy relationships between community centers and other nodes should be enlarged. In this section, we refine the fuzzy relations, which can make the center nodes stand out.

For each neighbor (e.g., X) of node V, there is a path from X to the node with greatest centrality while recursively finding the NGC nodes of X, NGC(X), NGC(NGC(X)), and so on. Let *frequent* refer to the number of such paths passing through node V. $r(V)$ is the rate of neighbors that can travel to node V.

$$r(V) = \frac{frequent}{deg(V)} \tag{6}$$

Thus, $r(V)$ indicates the proportion of the neighbors of node V that can travel to V while finding the NGC node recursively. A larger value of $r(V)$ means that V is more likely to be a community center.

Then, the fuzzy relation from V to NGC(V) can be refined with the following formula.

$$R^*(V, \mathrm{NGC}(V)) = \begin{cases} 1 - r(V), & R(V, \mathrm{NGC}(V)) < 0.5 \ \& \ r(V) < 0.5 \\ R(V, \mathrm{NGC}(V)), & others \end{cases} \tag{7}$$

where $R(V, \mathrm{NGC}(V))$ is the fuzzy relation without refinement. If the fuzzy relation $R(V, \mathrm{NGC}(V))$ is no less than 0.5, V is more likely to be a non-center node, whose fuzzy relation to the NGC node will not be refined. In other words, if the fuzzy relation between node V and its NGC node is no less than 0.5, V cannot be a community center since it is tightly associated with its NGC node.

If the fuzzy relation $R(V, \mathrm{NGC}(V))$ is small, node V may be a potential community center. However, some pseudo center nodes may be mixed in, whose fuzzy relations should be refined. Under the premise of $R(V, \mathrm{NGC}(V))$ being less than 0.5, $r(V)$ is used to identify the pseudo center nodes: if $r(V)$ is less than 0.5, V is more likely to be a pseudo center node,

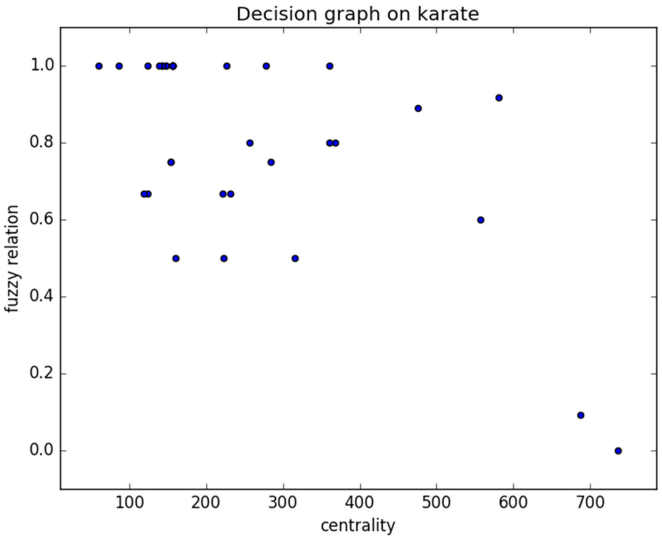


FIGURE 4. **The decision graph of the network *karate*.**

and the fuzzy relation between V and NGC(V) should be revised to $1 - r(V)$. In this manner, the gap between non-center nodes and center nodes will be conspicuous.

F. DECISION GRAPH

We first calculate the centrality of each node and the refined fuzzy relation from any node to its NGC node. After that, a so-called decision graph is generated, whose horizontal axis is centrality and who vertical axis is each node's refined fuzzy relation to its NGC node. Figure 4 is the decision graph of the network *karate* [35], which actually contains two communities.

It can be concluded from the decision graph that the nodes located in the bottom right corner, which are relatively far from other nodes, are extraordinary. In fact, they are the centers of communities. *A* and *B* are two distinct outliers. They are two core nodes, which are the beginnings of two communities, respectively. The effect of the decision graph on the construction of the community structure is crucial.

G. CONSTRUCT THE COMMUNITY STRUCTURE

The community structure can be derived from the decision graph, and Algorithm 2 shows the main procedure. First, nodes in the network are sorted in descending order based on centrality. *comnumber* refers to the community number, whose initial value is 0. It is noted that the first community is numbered to 1. Then, each node that has not been affiliated with any community (represented by V) is fetched sequentially from the ordered nodes list. Let $R^*(V, \mathrm{NGC}(V))$ denote the fuzzy relation from V to its NGC node. If $R^*(V, \mathrm{NGC}(V))$ is smaller than the threshold *delta*, which means that V is a new community center, a new community is created. Otherwise, V will be marked with the community number of its NGC node.

Figure 5 shows the community structure recognized by CDFR for the network *karate* when *delta* is 0.40. *karate* is divided into 2 communities (represented by blue and red, respectively), which is identical to the true community structure.

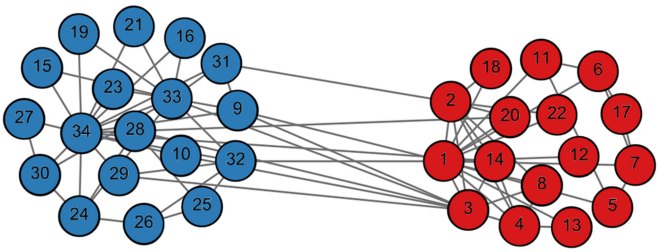


FIGURE 5. **The community structure recognized by CDFR in the network *karate* (delta = 0.40).**

Algorithm 2. Construct Community Structure.

Input: Network: G
Threshold: *delta*
Output: Communities: $\{C_1, C_2, \ldots, C_{communber}\}$

1. $G' \leftarrow$ sort the nodes of G in descending order based on centrality
2. *comnumber* $\leftarrow 0$
3. **For** each V in G' **do**:
4. **If** $R^*(V, \mathrm{NGC}(V)) < delta$ **then**:
   //Create a new community
5. $comnumber \leftarrow comnumber + 1$
6. $C_{comnumber} \leftarrow \{V\}$
7. **Else**:
8. $k \leftarrow$ the community number of NGC(V)
   //Add V into the community to which NGC(V) belongs
9. $C_k \leftarrow C_k \cup \{V\}$
10. **End If**
11. **End For**
12. **Return** $\{C_1, C_2, \ldots, C_{communber}\}$

How is the threshold *delta* determined? We can select the appropriate threshold *delta* by observing the decision graph. In general, there is a relatively conspicuous gap between the community' central nodes and non-central nodes, so we can easily choose an appropriate *delta* in the gap. Therefore, the nodes below the threshold *delta* are the community central nodes, and the other nodes are the non-central nodes.

Alternately, when the gap between central nodes of the communities and non-central nodes in the decision graph is not conspicuous, we need to make a rough estimation with respect to the number of communities. Suppose that the network will be divided into *comnumber* communities. Set a threshold value (which is called *delta* in this paper) with respect to the vertical axis in the corresponding decision graph so that the number of nodes whose fuzzy relations to their NGC nodes are smaller than that threshold is approximately *comnumber*.

H. TIME COMPLEXITY

Assume that the average degree of each node in the network is d and that the average hop count from a node to its NGC node is L. The average time complexity of computing the fuzzy relation is $\mathrm{O}(n^* d^L)$. In fact, the *NGC* nodes are often

not too far away, and L is relatively small. In addition, the time complexity of sorting is $O(n^* \log n)$.

IV. EXPERIMENT

In this part, the effectiveness of our CDFR algorithm will be exhibited through several experiments. First, the classical network datasets used in this paper will be presented. Second, we will introduce several evaluation criteria with respect to the performance of our algorithm. Third, the experimental results of CDFR on these data sets will be shown in detail, and comparisons between CDFR and other related algorithms will be given. Finally, experiments of CDFR on large networks will be carried out.

A. CLASSICAL DATA SETS

In this section, some classical network data sets used in our experiment will be introduced.

The *karate* data set [35] is a network that was derived from the United States Karate Club, which has 34 members. Each node represents a member, and an edge indicates that the two corresponding members are friends. This network was proposed by Zachary, and it contains two communities.

The *strike* data set [36], [37] refers to a social network in a wood processing factory. The new managers changed the benefits of employees, which did not make the employees satisfied, and they went on strike. The network contains several communities that had a relation to the strike. In fact, the network can be divided into 3 communities (Spanish-speaking, young English-speaking and old English-speaking) based on ethnicity and age.

The *dolphins* network [38] contains 62 dolphins living in New Zealand. Lusseau *et al.* found that the interaction relations between the dolphins formed a social network, where the nodes are dolphins and the edges link dolphins that make contact frequently. This network actually contains two communities.

The *US politics books* data set [39] is a network regarding *US politics books*, which was sold on Amazon. Nodes in the network refer to books, and the edge between any two books indicates that they were bought at the same time frequently. This network comprises 2 communities.

The *football* data set [12] refers to the network of American football games between Division IA colleges during the regular season of Fall 2000. Nodes in the network refer to the colleges, and an edge between two nodes means that the two colleges played a game. This network has 12 communities.

B. EVALUATIONS

There are many evaluation criteria regarding the performance of community detection algorithms, and they can be divided into two categories: the first is with ground-truth (e.g., ARI [40], [41], NMI [42] and purity [43]), and the second is without ground-truth (e.g., modularity Q [18]).

1) evaluation metrics with ground-truth

The ARI is one of the evaluation metrics with ground-truth, and it can be calculated with the following formula:

$$ARI = \frac{\sum_{i,j}\binom{m_{ij}}{2} - \frac{\sum_i \binom{m_{i.}}{2}\sum_j \binom{m_{.j}}{2}}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{m_{i.}}{2} + \sum_j \binom{m_{.j}}{2}\right] - \frac{\sum_i \binom{m_{i.}}{2}\sum_j \binom{m_{.j}}{2}}{\binom{N}{2}}} \tag{8}$$

where m_{ij} is the value of the ith row and jth column of the contingency matrix [44]. $m_{i.}$ is the sum of the ith row, and m_{j} is the sum of the jth column of the contingency matrix. N is the number of nodes in the network.

Another evaluation is NMI, which was introduced in [42].

$$NMI = \frac{-2\sum_{i,j} m_{ij} \cdot \log\left(\frac{m_{ij}N}{m_{i.}m_{.j}}\right)}{\sum_i m_{i.}\log\left(\frac{m_{i.}}{N}\right) + \sum_j m_{.j}\log\left(\frac{m_{.j}}{N}\right)} \tag{9}$$

The purity is the third metric with ground-truth used in this paper.

$$purity(R, C) = \frac{1}{N} \cdot \sum_i \max_j \left|R_i \cap C_j\right| \tag{10}$$

where $R = \{R_1, R_2, \ldots, R_{k'}\}$ is the partition of the network, which is generated by our CDFR algorithm. $C = \{C_1, C_2, \ldots, C_k\}$ is the true community structure of the network.

2) evaluation metrics without ground-truth

The most commonly used evaluation of the result of a community detection algorithm is modularity Q.

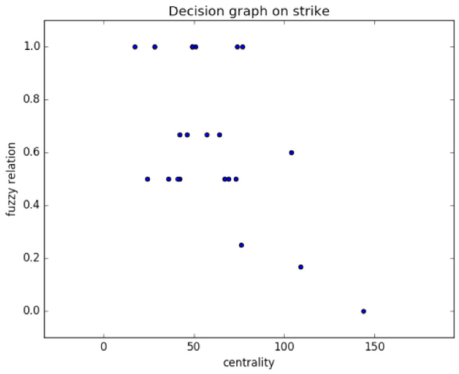
$$Q = \sum_{c \in \{C_1, C_2, \ldots, C_k\}} \left(\frac{e_c}{E} - \left(\frac{d_c}{2E}\right)^2\right) \tag{11}$$

where $\{C_1, C_2, \ldots, C_k\}$ is the community structure recognized by our algorithm and E is the number of edges of the network. In addition, e_c is the number of internal edges of the community c, and d_c is the sum of the degrees of nodes in the community c.

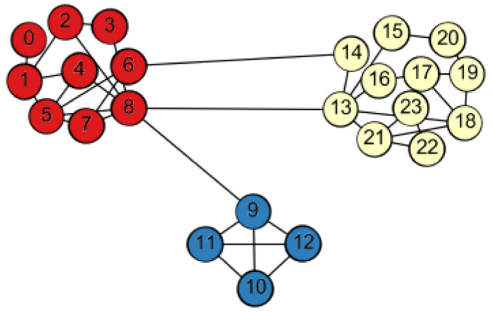
C. EXPERIMENTAL RESULTS AND COMPARISONS

In many cases, the gap between the community center nodes and non-center nodes is relatively obvious. Thus, we can choose a suitable threshold by observing the decision graph. A node whose fuzzy relation is less than the threshold is a community center node, and otherwise it is a non-center node.

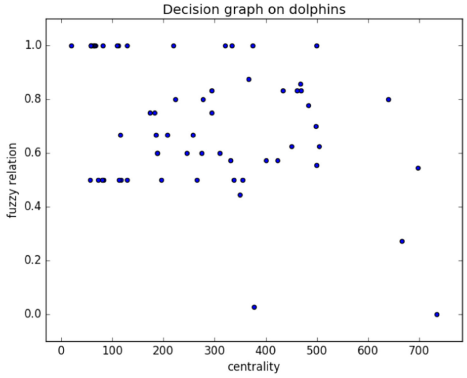
Figure 6. shows the decision graphs and experimental results of CDFR. (a), (c), (e) and (g) refer to the decision graphs of CDFR on the networks *strike*, *dolphins*, *US politics books* and *football*, respectively. (b), (d), (f) and (h) refer to the communities recognized by CDFR in the networks *strike*, *dolphins*, *US politics books* and *football*, respectively.



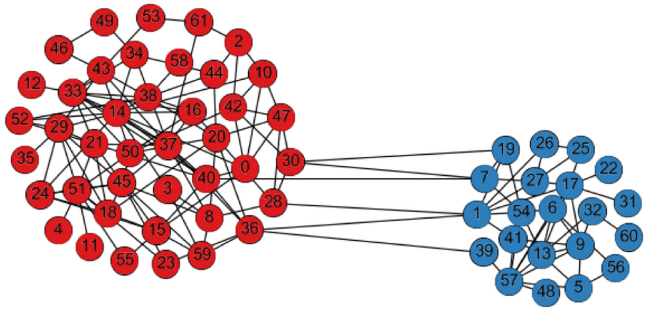
(a) The decision graph of *strike*.



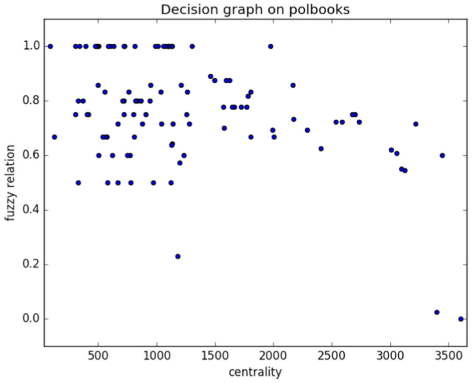
(b) The communities of *strike* obtained by CDFR while *delta*=0.4.



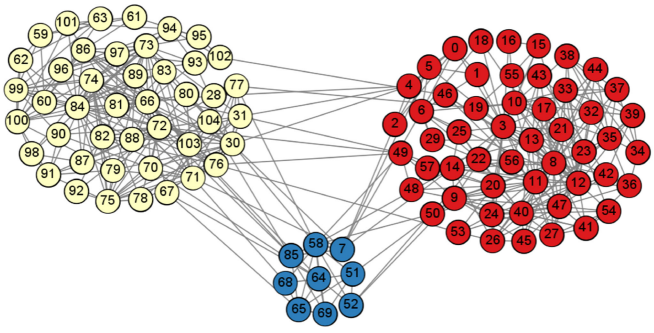
(c) The decision graph of *dolphins*.



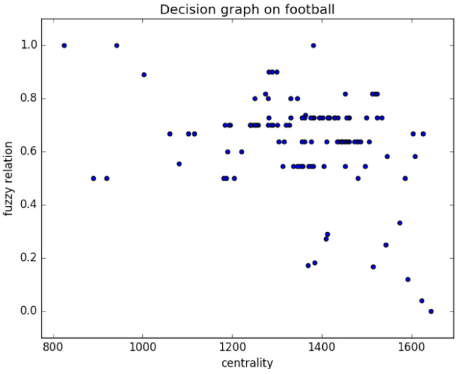
(d) The communities of *dolphins* recognized by CDFR while *delta*=0.2.



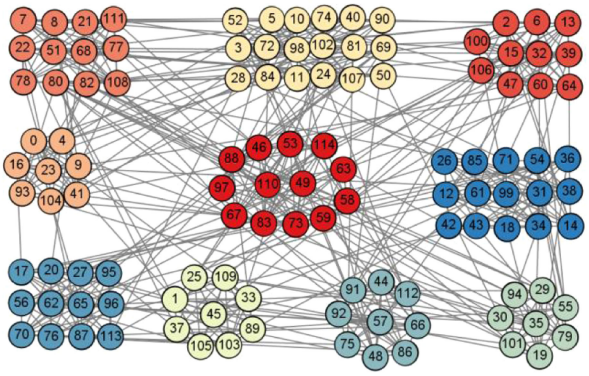
(e) The decision graph of *US politics books*.



(f) The communities of *US politics books* recognized by CDFR while *delta*=0.4.



(g) The decision graph of *football*



(h) The communities of *football* recognized by CDFR while *delta*=0.4

FIGURE 6. The decision graphs and the communities recognized by CDFR in the networks *strike*, *dolphins*, *US politics books* and *football*, respectively.

TABLE 2. A comparison of various community detection algorithms.

Dataset	Algorithm	K'	ARI	purity	NMI	Q
karate	FastModularity	3	0.680	0.970	-	0.380
	cFinder	3	0.705	0.065	-	0.182
	SCAN	4	0.314	0.764	-	0.312
	TopLeader	2	**1.000**	**1.000**	-	0.371
	Newman	5	0.46	-	0.57	0.40
	Louvain	4	0.46	-	0.58	**0.41**
	Walktrap	5	0.33	-	0.50	0.35
	LICOD	3	0.62	-	0.60	0.24
	CDFR (*delta* = 0.40)	3	**1.000**	**1.000**	**1.000**	0.371
strike	FastModularity	4	0.664	0.958	-	**0.555**
	cFinder	6	0.348	**1.000**	-	0.485
	SCAN	3	0.848	0.958	-	0.547
	TopLeader	3	**1.000**	**1.000**	-	0.548
	CDFR (*delta* = 0.40)	3	**1.000**	**1.000**	**1.000**	0.548
dolphins	Newman	5	0.39	-	0.55	**0.51**
	Louvain	5	0.32	-	0.51	**0.51**
	Walktrap	4	0.41	-	0.53	0.48
	LICOD	2	0.32	-	0.41	0.35
	CDFR (*delta* = 0.20)	2	**0.935**	**0.984**	**0.889**	0.379
US politics books	Newman	5	**0.68**	-	0.55	0.51
	Louvain	4	0.55	-	0.57	**0.52**
	Walktrap	4	0.65	-	0.53	0.50
	LICOD	6	0.67	-	**0.68**	0.42
	CDFR (*delta* = 0.40)	3	0.678	**0.867**	0.567	0.491
football	Newman	10	0.77	-	0.87	0.59
	Louvain	10	0.80	-	0.89	**0.60**
	Walktrap	10	**0.81**	-	0.83	**0.60**
	LICOD	16	0.69	-	0.83	0.49
	CDFR (*delta* = 0.40)	10	0.809	**0.930**	**0.899**	0.591

In the decision graph of CDFR for *strike* (i.e., Figure 6a), it is evident that three points in the lower right corner are conspicuous, which are the centers of three communities. Suppose that the threshold *delta* of the fuzzy relation is 0.40; then, we get the three communities of the network *strike* (i.e., Figure 6b). For the same reason, the community structures of *dolphins*, *US politics books* and *football* can be obtained by use of the corresponding decision graphs.

Comparisons among FastModularity [20], CFinder [45], SCAN [46], TopLeader [28], Newman [18], Louvain [21], Walktrap [47], LICOD [29] and CDFR are shown in Table II. "-" indicates that relevant data has not been found, and we only list the data results we have found. *delta* is the threshold of the fuzzy relation, and K' is the community number of the network recognized by our algorithm or a parameter in other algorithms. In addition, the values with two decimal places are derived from [29], and the values with three decimal places are taken from [28].

In the networks of *karate* and *strike*, if the threshold *delta* of the fuzzy relation is 0.40, the results of CDFR on *karate* and *strike* are optimal with respect to ARI, purity, and NMI. Although CDFR is not dominant in modularity Q, it can recognize the real community structure as compared to the ground-truth.

In the decision graph of CDFR on the network *dolphins*, which actually contains 2 groups, two isolated nodes located at the bottom are different from other nodes and may be the community centers. When *delta* assumes the value 0.20, two communities are formed. In this case, CDFR performs best with respect to ARI, Purity and NMI. However, the modularity Q is only 0.379. Alternately, delta may take the value 0.40, and as a result, three community centers are selected; for more information regarding the selection of *delta*, please refer to Section V.

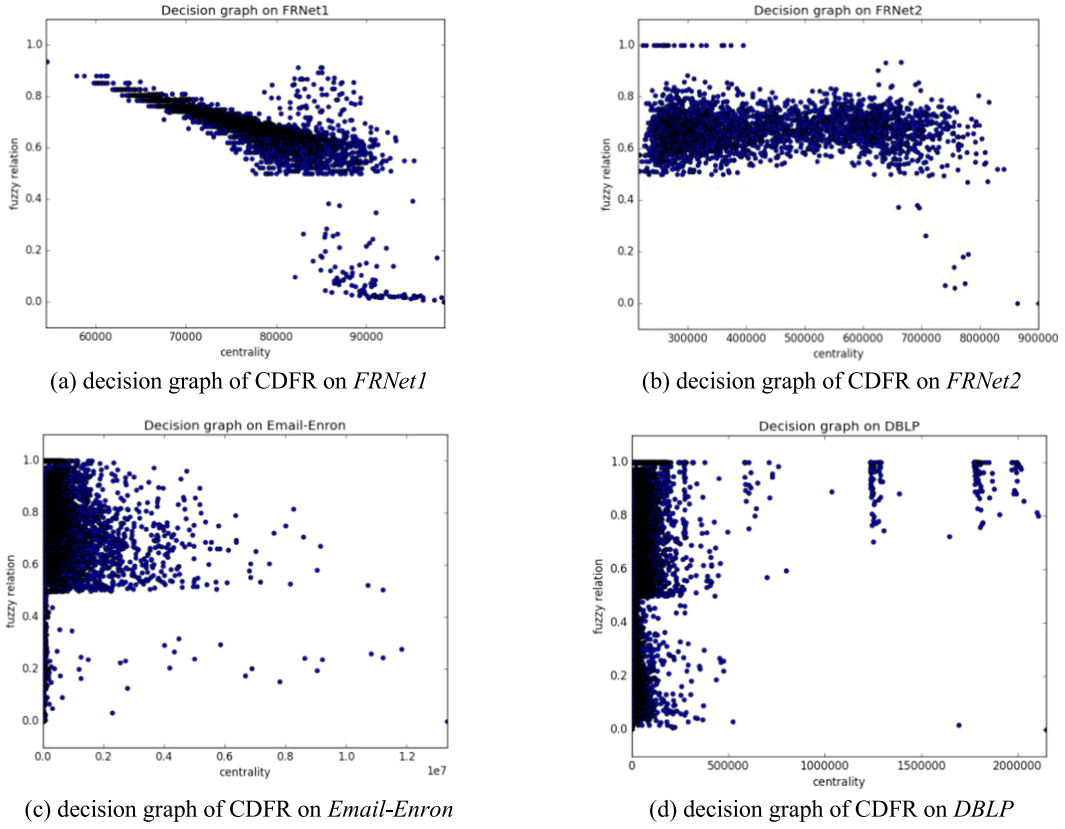
In the decision graph of CDFR on *US politics books*, 3 points below *delta* = 0.40 are special and independent, and thus they are regarded as community centers. In the network of *US politics books*, the algorithm Newman exhibits the optimal value for the ARI metric. CDFR has the best performance in Purity and NMI.

There is an obvious gap approximately 0.40 in the decision graph of the network *football*. When *delta* is 0.40, CDFR is the best with respect to NMI when compared to Louvain, Walktrap and LICOD. However, Louvain and Walktrap obtain the largest modularity Q.

D. EXPERIMENTS ON LARGE NETWORKS

To verify the performance of CDFR on large-scale networks, two large-scale artificial networks and two real-world networks are used in our experiments.

The implementation of the Newman_Watts_Strogatz method [48] in package networks is used to generate an artificial network *FRNet1*, which contains 5,000 nodes and 104,997 edges. Another artificial network *FRNet2* is generated



(a) decision graph of CDFR on *FRNet1*

(b) decision graph of CDFR on *FRNet2*

(c) decision graph of CDFR on *Email-Enron*

(d) decision graph of CDFR on *DBLP*

FIGURE 7. Decision graphs of CDFR on large networks: *FRNet1*, *FRNet2*, *Email-Enron*, and *DBLP*.

by the Geometric Preferential Attachment model [49] in the SNAP package [50], and it consists of 3,000 nodes and 104,633 edges. In addition, the two large-scale real-world networks *Email-Enron* [51], [52] and *DBLP* [53] are used in the experiment, which are excerpted from the SNAP data set. *Email-Enron* contains 36,692 nodes and 183,831 edges, and *DBLP* contains 317,080 nodes and 1,049,866 edges.

In addition, these networks do not have ground-truths, or their ground-truths are overlapping communities, so the evaluation metrics for non-overlapping ground-truths (e.g., NMI, purity and ARI) can no longer be used. In this part, only the modularity Q is used to evaluate the community detection results. Because LPA has low time complexity suitable for the analysis of large networks, we compare the performance of LPA and CDFR on these 4 relatively large-scale networks. The implementation of LPA that we used is from the *igraph* package. Because the results of the LPA algorithm could be different in each run, we take the average values and standard deviations of 100 runs.

Figure 7 refers to the decision graphs of CDFR on the networks *FRNet1*, *FRNet2*, *Email-Enron* and *DBLP*. As shown in the figure, the gap between the community center nodes and non-central nodes is relatively obvious. Thus, the thresholds (i.e., values of *delta*) for *FRNet1*, *FRNet2*, *Email-Enron* and *DBLP* are all set to 0.4. The experimental results (i.e., Q) are given in Table III, which shows that CDFR is better than the LPA algorithm on these large-scale networks.

V. DISCUSSION

In this section, some aspects of this paper should be discussed in depth. First, we discuss the selection of the threshold of the fuzzy relation (i.e., *delta*). Second, the impact of the unrefined fuzzy relation on the performance will be discussed. Third, the central nodes of communities will be discussed.

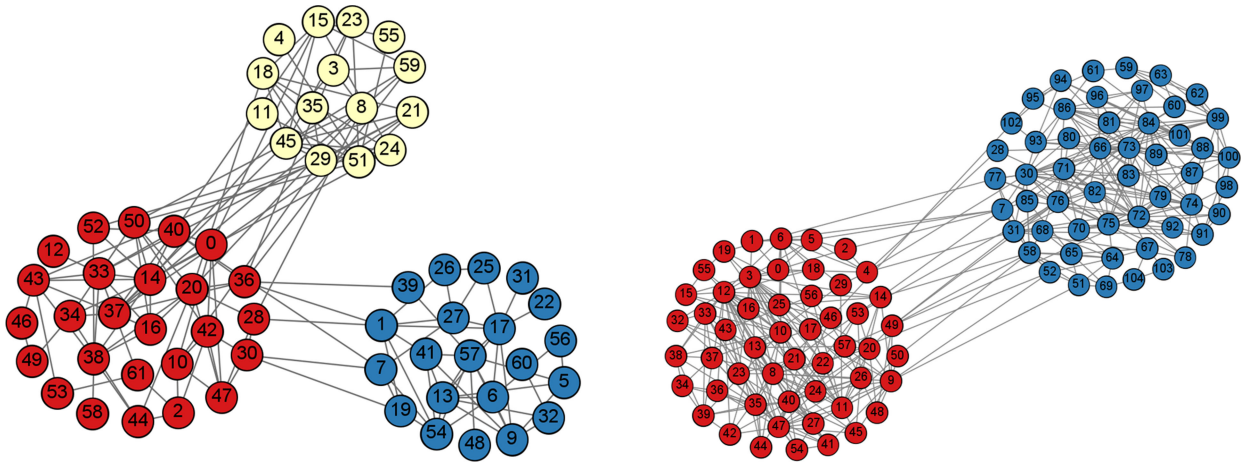
A. ON THE THRESHOLD OF FUZZY RELATION

In the previous work, we determined the threshold *delta* by observing the decision graph.

In many cases (e.g., *karate*, *strike*, *dolphins*, *US politics books* and *football*), because there is a gap between the community center nodes and other non-center nodes in decision graphs, the threshold *delta* is easy to select, and these results are not given here.

TABLE 3. The experimental results of CDFR on large network data sets.

Algorithms	*FRNet1*	*FRNet2*	*Email-Enron*	*DBLP*
LPA	0.570 ± 0.009	0.545 ± 0.155	0.313 ± 0.041	0.683 ± 0.012
CDFR	**0.585**	**0.571**	**0.511**	**0.727**



(a) CDFR on *dolphins* (*delta*=0.40) (b) CDFR on *US politics books* (*delta*=0.15)

FIGURE 8. The communities obtained by CDFR on the networks *dolphins* and *US politics book* when *delta* assumes different values.

However, for some networks, the gap of the center nodes and non-center nodes is not conspicuous in the decision graph (e.g., *dolphins* and *US politics books*). The threshold *delta* may have several choices with respect to the networks *dolphins* and *US politics books*. In our previous work, *delta* takes the value 0.20 on *dolphins* to form two communities. However, another choice (e.g., $delta = 0.40$) may yield different results.

Figure 8a shows the community structure of *dolphins*, which is divided into 3 communities when *delta* is 0.40. Similarly, Figure 8b shows the communities of *US politics books* recognized by CDFR when $delta = 0.15$. The evaluation metrics of CDFR with different *delta* values on *dolphins* and *US politics books* are given in Table IV. For the network *dolphins*, when $delta = 0.40$, CDFR can divide it into 3 communities. Although there are some differences between the obtained community structure (i.e., CDFR with $delta = 0.40$) and the real communities, the modularity Q is higher than CDFR with $delta = 0.20$. For the network *US politics books*, the purity and NMI of CDFR with $delta = 0.15$ is higher, and the ARI and Q of CDFR with $delta = 0.40$ are larger.

B. ON THE UNREFINED FUZZY RELATION

Consider that the unrefined fuzzy relation is regarded as the fuzzy relation directly, and the corresponding algorithm is called CDFR-U. That is to say, in CDFR-U, $R(V, \mathrm{NGC}(V))$ is adopted in Algorithm 2, instead of $R^*(V, \mathrm{NGC}(V))$. Figure 9 shows the decision graphs of *karate*, *strike*, *dolphins*, *US politics books* and *football* for CDFR-U.

There are no significant changes between CDFR and CDFR-U on *karate* and *strike* with respect to the decision graphs. Thus, the communities recognized by CDFR ($delta = 0.40$) and CDFR-U ($delta = 0.40$) are identical. Alternately, although CDFR-U generates a changed decision graph for *dolphins*, the experimental results of CDFR-U with $delta = 0.15$ are identical to those of CDFR with $delta = 0.20$ on *dolphins*. Thus, we only list the decision graphs of CDFR-U on *karate*, *strike* and *dolphins*, and their community structures are not given here anymore.

However, there is no obvious gap between community central nodes and other non-central nodes in the decision graphs of CDFR-U on *US politics books* and *football*. Thus, it is not easy to find a suitable threshold *delta* because there are many options for *delta*. Different *delta* values will lead to different community structures being identified by our algorithm.

Figure 10 shows the community structures of *US politics books* and *football* recognized by CDFR-U when *delta* takes different values. The evaluation metrics of CDFR-U with different *delta* values on *US politics books* and *football* are given in Table V.

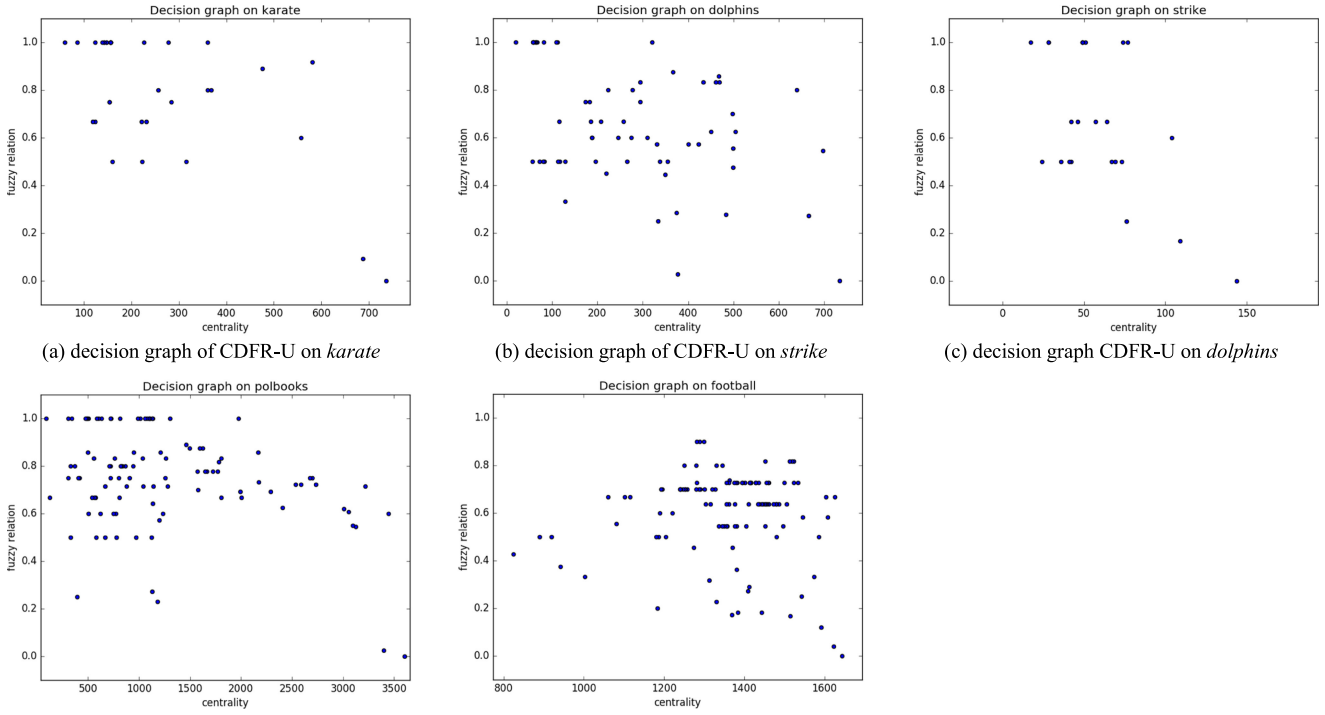
From Figure 10 and Table V, it can be observed that CDFR-U with a larger *delta* will obtain more communities. CDFR-U can discover more hub nodes or small-scale communities when *delta* takes larger values.

C. ON THE CENTRAL NODES

In CDFR, for each node of the network, if its fuzzy relation is less than the threshold *delta*, it will be considered the central

TABLE 4. The experimental results of CDFR on *Dolphins* and *US politics books* with different values of *delta*.

networks	*delta*	K'	ARI	Purity	NMI	Q
dolphins	0.20	2	**0.935**	**0.984**	**0.889**	0.379
	0.40	3	0.540	0.742	0.662	**0.491**
US politics books	0.15	2	0.667	**0.914**	**0.598**	0.457
	0.40	3	**0.678**	0.867	0.567	**0.491**



(a) decision graph of CDFR-U on *karate* (b) decision graph of CDFR-U on *strike* (c) decision graph CDFR-U on *dolphins*

(d) decision graph of CDFR-U on *US politics books* (e) decision graph of CDFR-U on *football*.

FIGURE 9. The decision graphs of CDFR-U on *karate*, *strike*, *dolphins*, *US politics books*, and *football*.

TABLE 5. The experimental results of CDFR-U on *US politics books* and *football* with different values of *delta*.

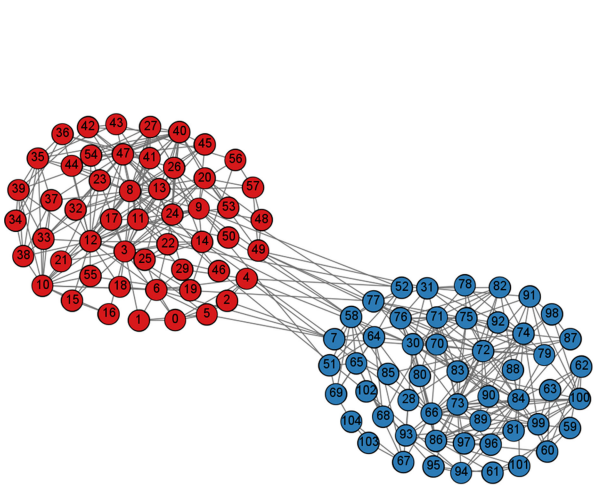
network	*delta*	K'	ARI	Purity	NMI	Q
US politics books	0.15	2	**0.667**	**0.914**	**0.598**	0.457
	0.27	4	0.662	0.857	0.555	0.486
	0.40	5	0.654	0.810	0.538	**0.504**
football	0.24	9	0.625	**0.922**	0.832	0.553
	0.30	12	0.821	0.913	0.902	**0.580**
	0.35	15	**0.845**	0.861	**0.911**	0.556

node of a community. The value of the parameter *delta* could be set by observing the decision graph. This allows us to discover the communities with central nodes with large centralities, as well as those with small centralities. This is because the centralities of the central nodes could vary significantly in various real social networks.

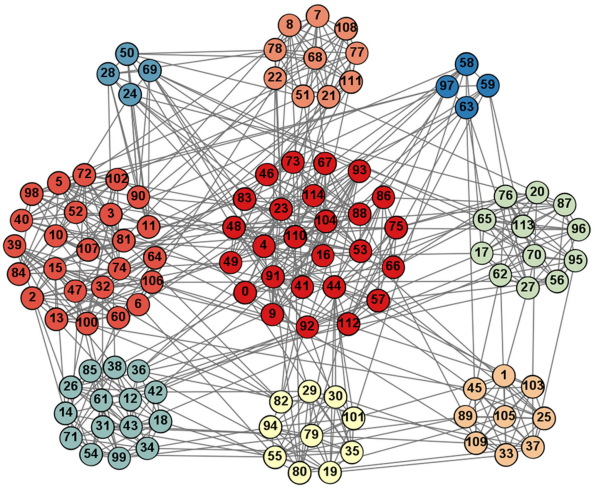
However, we could also define a central node as follows: its fuzzy relation is less than the threshold *delta* and its centrality is larger than a threshold *beta*. Under such conditions, some communities whose central nodes have small centralities (no larger than *beta*) cannot be found. This is a good choice when we do not want to detect communities for which their central nodes have relatively small centrality.

VI. CONCLUSION

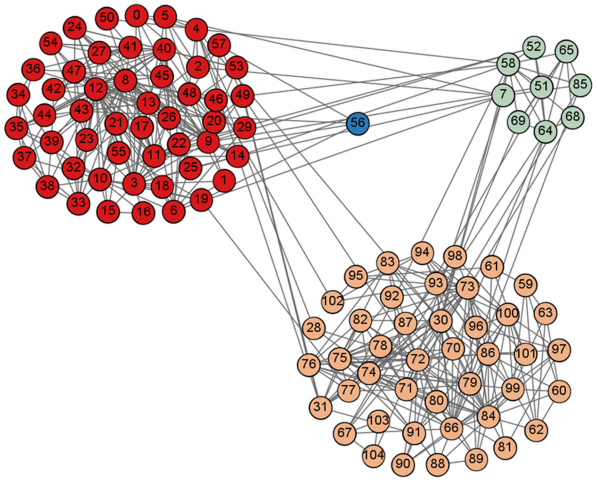
Non-overlapping community detection is a very significant research topic in the social networks analysis and mining field. The goal of non-overlapping community detection is to divide the network into several communities, and each node belongs to only one community. Thus far, many non-overlapping community detection algorithms have been proposed. The role of the centrality of a node has been widely studied in non-overlapping community detection, as well as the relation between a node and its neighbors. However, in existing algorithms, the importance of the Nearest node with Greater Centrality has not been paid much attention, and the same could be said for the fuzzy relation between a node and its Nearest node with Greater Centrality. In this paper, a community detection named CDFR, which is based on fuzzy relations, is proposed to discover non-overlapping communities. First, the centrality based on the micro-environment is calculated, and the so-called NGC node is introduced. Second, the fuzzy relation from a node to its NGC node is calculated based on the composition of fuzzy relations. Finally, each node in the network is affiliated with the community that its NGC node belongs to if the fuzzy relation is large enough. In particular, a node becomes the center node of a new community if the fuzzy relation is smaller than a threshold *delta*. In addition, CDFR is compared to various



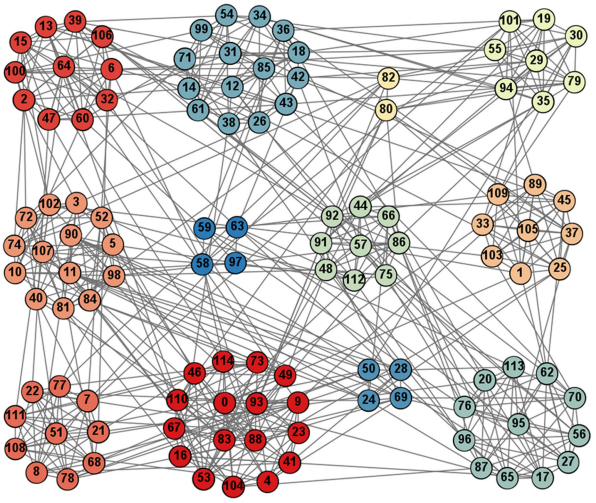
(a) communities of *US politics books* recognized by CDFR-U (*delta*=0.15)



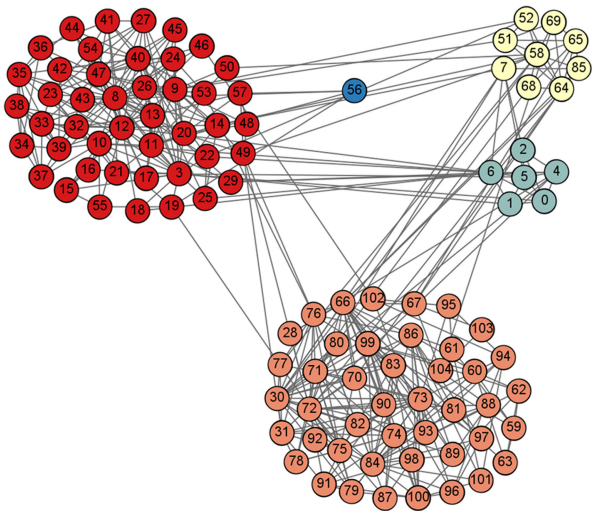
(b) communities of *football* recognized by CDFR-U (*delta*=0.24)



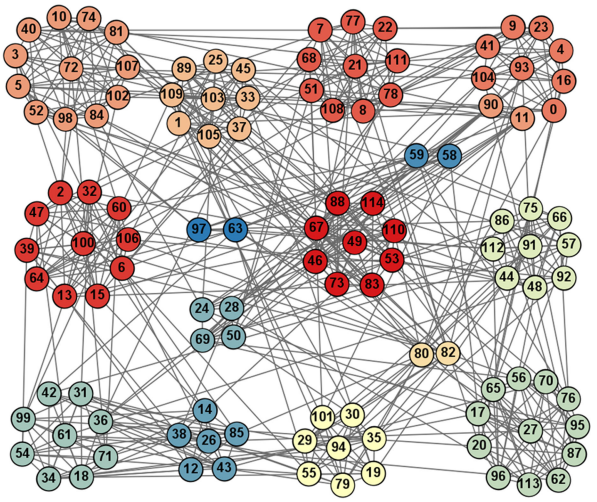
(c) communities of *US politics books* recognized by CDFR-U (*delta*=0.27)



(d) communities of *football* recognized by CDFR-U (*delta*=0.30)



(e)communities of *US politics books* recognized by CDFR-U(*delta*=0.40)



(f) communities of *football* recognized by CDFR-U (*delta*=0.35)

FIGURE 10. The communities obtained by CDFR-U on *US politics books* and *football* when *delta* takes different values: (a), (c) and (e) refer to the community structures of *US politics books* recognized by CDFR-U when $delta = 0.15$, $delta = 0.27$, and $delta = 0.40$, respectively; (b), (d) and (f) are the community structures of *football* recognized by CDFR-U when $delta = 0.24$, $delta = 0.30$, and $delta = 0.35$, respectively.

classical community detection algorithms, and the results verify the effectiveness of our algorithm.

In future work, the centrality will be discussed in depth. For example, the edge centrality may be used to calculate the fuzzy relation instead of the one-way intersection rate. In addition, an improved version of this algorithm that can be used to detect local community, overlapping community and dynamic community structures will be explored in the future.

ACKNOWLEDGMENTS

This work is partially supported by Anhui Provincial Natural Science Foundation (No. 1408085MKL07).

REFERENCES

[1] S. Gregory, "Fuzzy overlapping communities in networks," *J. Statist. Mech. Theory Experiment*, vol. 2011, 2011, Art. no. P02017.

[2] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdisciplinary Rev. Data Mining Knowl. Discovery*, vol. 6, pp. 115–135, 2016.

[3] S. Fortunato, "Community detection in graphs," *Physics Rep.*, vol. 486, pp. 75–174, 2010.

[4] M. Plantié and M. Crampes, "Survey on social community detection," in *Social Media Retrieval*, N. Ramzan, R. van Zwol, J. Lee, K. Clüver, and X. Hua, Eds. Berlin, Germany: Springer, 2013, pp. 65–85.

[5] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Sci.*, vol. 344, pp. 1492–1496, 2014.

[6] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering," *Physica A Statist. Mech. Appl.*, vol. 374, pp. 483–490, 2007.

[7] P. G. Sun, "Community detection by fuzzy clustering," *Physica A: Statist. Mech. Appl.*, vol. 419, pp. 408–416, 2015.

[8] P. G. Sun, L. Gao, and S. S. Han, "Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks," *Inf. Sci.*, vol. 181, pp. 1060–1071, 2011.

[9] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Syst. Techn. J.*, vol. 49, pp. 291–307, 1970.

[10] A. Pothen, "Graph partitioning algorithms with applications to scientific computing," in *Parallel Numerical Algorithms*, D. E. Keyes, A. Sameh, and V. Venkatakrishnan, Eds. Berlin, Germany: Springer, 1997, pp. 323–368.

[11] E. R. Barnes, "An algorithm for partitioning the nodes of a graph," *SIAM J. Algebraic Discrete Methods*, vol. 3, pp. 541–550, 1982.

[12] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proc. Natl. Academy Sci. United Nation America*, vol. 99, pp. 7821–7826, 2002.

[13] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinf.*, vol. 22, pp. 2283–2290, 2006.

[14] S. Moon, J.-G. Lee, M. Kang, M. Choy, and J.-W. Lee, "Parallel community detection on large graphs with MapReduce and GraphChi," *Data Knowl. Eng.*, vol. 104, pp. 17–31, 2016.

[15] J. W. Pinney and D. R. Westhead, "Betweenness-based decomposition methods for social and biological networks," in *Interdisciplinary Statistics and Bioinformatics*, S. Barber, P. Baxter, K. Mardia, and R. Walls Eds. Leeds, U.K.: Leeds University Press, 2006, pp. 87–90.

[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proc. Natl. Academy Sci. United States America*, vol. 101, pp. 2658–2663, 2004.

[17] M. J. Rattigan, M. Maier, and D. Jensen, "Graph clustering with network structure indices," in *Proc. 24th Int. Conf. Mach. Learning*, 2007, pp. 783–790.

[18] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Rev. E*, vol. 69, 2004, Art. no. 026113.

[19] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Rev. E*, vol. 69, 2004, Art. no. 066133.

[20] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Rev. E*, vol. 70, 2004, Art. no. 066111.

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech.: Theory Experiment*, vol. 2008, 2008, Art. no. P10008.

[22] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Rev. E*, vol. 76, 2007, Art. no. 036106.

[23] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2012, pp. 25–36.

[24] W. Hu, "Finding statistically significant communities in networks with weighted label propagation," *Soc. Netw.*, vol. 2, 2013, Art. no. 34013.

[25] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Physics*, vol. 12, 2010, Art. no. 103018.

[26] Z.-H. Wu, Y.-F. Lin, S. Gregory, H.-Y. Wan, and S.-F. Tian, "Balanced multi-label propagation for overlapping community detection in social networks," *J. Comput. Sci. Technol.*, vol. 27, pp. 468–479, 2012.

[27] T. You, B.-C. Shia, and Z.-Y. Zhang, "Community detection in complex networks using density-based clustering algorithm and manifold learning," *Phys. Stat. Mech. Appl.*, vol. 464, pp. 221–230, 2016.

[28] R. R. Khorasgani, J. Chen, and O. R. Zaïane, "Top leaders community detection approach in information networks," in *Proc. 4th SNA-KDD Workshop Social Netw. Mining Anal.*, 2010, pp. 228–235.

[29] Z. Yakoubi and R. Kanawati, "LICOD: A Leader-driven algorithm for community detection in complex networks," *Vietnam J. Comput. Sci.*, vol. 1, pp. 241–256, 2014.

[30] D. Shah and T. Zaman, "Community detection in networks: The leader-follower algorithm," *Proc. Workshop Net. Across Disciplines: Theory Appl.*, 2010, pp. 1–8.

[31] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[32] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Soc.*, vol. 25, pp. 163–177, 2001.

[33] L.-X. Wang, *A Course in Fuzzy Systems*. Upper Saddle River, NJ, USA: Prentice-Hall Press, 1999.

[34] M. Stachowicz and M. Kochanska, "Graphic interpretation of fuzzy sets and fuzzy relations," *Math. Service of Man*, A. Ballester, D. Cardins, E. Trillas, pp. 620–629, 1982.

[35] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropological Res.*, vol. 33, pp. 452–473, 1977.

[36] W. De Nooy, A. Mrvar, and V. Batagelj, *Exploratory Social Network Analysis with Pajek*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[37] J. H. Michael, "Labor dispute reconciliation in a forest products manufacturing facility," *Forest Products J.*, vol. 47, 1997, Art. no. 41.

[38] D. Lusseau, "The emergent properties of a dolphin social network," *Proc. Royal Soc. London B: Biological Sci.*, vol. 270, pp. S186–S188, 2003.

[39] V. Krebs, Political books network. (2004). [Online]. Available: http://www-personal.umich.edu/~mejn/netdata/polbooks.zip

[40] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Proc. Int. Conf. Artif. Neural Netw.*, 2009, pp. 175–184.

[41] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[42] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," *J. Mach. Learning Res.*, vol. 3, pp. 583–617, 2003.

[43] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[44] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 877–886.

[45] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.

[46] X. Xu, N. Yuruk, Z. Feng, and T. A. Schweiger, "Scan: A structural clustering algorithm for networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 824–833.

[47] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Proc. Int. Symp. Comput. Inf. Sci.*, 2005, pp. 284–293.

[48] M. E. Newman and D. J. Watts, "Renormalization group analysis of the small-world network model," *Physics Lett. A*, vol. 263, pp. 341–346, 1999.

[49] A. D. Flaxman, A. M. Frieze, and J. Vera, "A geometric preferential attachment model of networks," *Internet Math.*, vol. 3, pp. 187–205, 2006.

[50] J. Leskovec and A. Krevl, SNAP Datasets: Stanford *Large Network Dataset Collection*. (2014). [Online]. Available: http://snap.stanford.edu/data

[51] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, pp. 29–123, 2009.

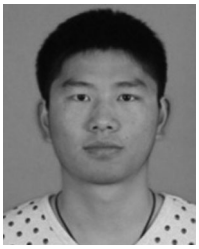
[52] B. Klimt and Y. Yang, Introducing the Enron Corpus. CEAS, 2004. [Online]. Available: http://www.cs.cmu.edu/~enron/

[53] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 745–754.

WENJIAN LUO (SM'15) received the BS and PhD degrees from Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China, in 1998 and 2003. He is presently an associate professor of the School of Computer Science and Technology, University of Science and Technology of China. His current research interests include machine learning and data mining, information security, computational intelligence, and applications. He is a member of the IEEE.

ZHENGLONG YAN received the BE degree from the College of Information Engineering, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shanxi, China, in 2013. He is currently working toward the master's degree at the School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China. His current research interests include data mining and applications.

CHENYANG BU received the BE degree from Hefei University of Technology, Hefei, China, in 2012. He is currently working toward the PhD degree in University of Science and Technology of China (USTC), Hefei, China. His research interests include evolutionary optimization and data mining engineering.



DAOFU ZHANG received the BE degree from Anhui University, Hefei, China, in 2015. He is currently working toward the MS degree in University of Science and Technology of China(USTC), Hefei, China. His research interests include machine learning and data mining.

