# An Evolutionary Multiobjective Optimization Based Fuzzy Method for Overlapping Community Detection

Ye Tian, Shangshang Yang, and Xingyi Zhang, *Senior Member, IEEE*

***Abstract*—In the last decade, the detection of overlapping communities has received increasing attention in network science. Among various clustering techniques, the fuzzy clustering has been widely adopted in overlapping community detection, since the soft assignment provided by it naturally meets the overlapping between multiple communities. The crucial step of fuzzy-clustering-based overlapping community detection is to find the optimal community centers, so that the overlapping communities can be obtained according to the membership degrees between nodes and community centers. In this article, we propose an evolutionary multiobjective optimization-based fuzzy method for overlapping community detection. In contrast to traditional fuzzy clustering methods, the proposed method optimizes the community centers by using a specially tailored multiobjective evolutionary algorithm. Moreover, it can also find an appropriate fuzzy threshold for each node, so that diverse overlapping community structures can be uncovered. In the experiments, we compare the proposed method with six state-of-the-art overlapping community detection approaches on synthetic and real-world networks with different scales and characteristics. The statistical results demonstrate that the proposed method can obtain the best results on most test instances.**

***Index Terms*—Complex network, evolutionary multiobjective optimization, fuzzy clustering, overlapping community detection.**

## I. Introduction

With the development of network science, complex networks have been applied in many fields to represent various kinds of complex systems, such as biological networks [1] and social networks [2]. Community structure denotes the division of a network into some groups, where the nodes within the same group have dense connections and those in different groups have sparse connections [3]. The detection of community structure has become an important technique in many fields, such as social science, network science, biological science, etc., [4].

Since the community detection problem was proposed by Girvan and Newman [3], there have been many approaches for uncovering nonoverlapping communities that each node belongs to one single community [5], [6]. However, the communities in many real-world networks are often overlapped with each other [7]. For example, one person is often a member of several social groups including clubs, friends, family, and colleagues, and a scientist may be active in several areas. In 2005, Palla *et al.* [7] extended the nonoverlapping community detection approach to uncover overlapping communities, and a variety of techniques have been adopted to detect overlapping communities from then on, including clique percolation theory [7], local expansion [8], label propagation [9], link community detection [10], fuzzy clustering [11], and evolutionary computation [12].

The fuzzy clustering has become an effective technique for overlapping community detection, since the fuzzy clustering that allows a node to belong to more than one cluster naturally meets the overlapping between multiple clusters [13]. The crucial step of fuzzy-clustering-based overlapping community detection is to find the optimal community centers. Once the community centers are determined, the membership matrix $U$ between nodes and community centers can be calculated, and the overlapping communities can be obtained according to $U$ and a predefined threshold $\lambda$. Following this idea, a number of fuzzy-clustering-based overlapping community detection approaches have been proposed in the last decade. For instance, Zhang *et al.* [14] proposed an overlapping community detection approach based on the combination of spectral mapping, fuzzy $c$-means (FCM) [15], and optimization of a quality function; Nepusz *et al.* [16] regarded the fuzzy-clustering-based overlapping community detection as a nonlinearly constrained optimization problem and solved it by a quadratic-time algorithm; and Wang *et al.* [17] proposed a fuzzy clustering algorithm based on local random walk and a new distance measure.

Although the abovementioned approaches have successfully employed fuzzy clustering to uncover overlapping communities, the fuzzy clustering is still criticized for the high sensitivity to community center initialization, which makes these approaches easily trapped into local optimum [18]–[20]. Besides, some parameters including the number of communities $k$ and the fuzzy threshold $\lambda$ should be predefined in these approaches, which are pivotal to the clustering result but difficult to be determined in advance [11], [21]–[23]. To address these issues, this article

Manuscript received February 27, 2019; revised May 22, 2019; accepted September 20, 2019. Date of publication October 2, 2019; date of current version October 30, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61672033, Grant 61822301, Grant 61876123, Grant 61906001, and Grant U1804262, in part by the State Key Laboratory of Synthetical Automation for Process Industries under Grant PAL-N201805, in part by the Anhui Provincial Natural Science Foundation under Grant 1808085J06 and Grant1908085QF271, and in part by the Recruitment Program for Leading Talent Team of Anhui Province (2019-16). *(Corresponding author: Xingyi Zhang.)*

Y. Tian is with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China (e-mail: field910921@gmail.com).

S. Yang and X. Zhang are with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: yangshang0308@gmail.com; xyzhanghust@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2019.2945241

proposes a fuzzy method for overlapping community detection, which uses a multiobjective evolutionary algorithm (MOEA) [24] to automatically optimize the community centers as well as the parameters in fuzzy clustering. The contributions of this article are as follows.

1) An evolutionary multiobjective optimization-based fuzzy method (EMOFM) is proposed for overlapping community detection. The proposed method adopts the idea of fuzzy $c$-medoids [25], where each community center is a node in the network (termed central node). The central nodes are optimized by a well-tailored MOEA, and the number of communities can also be automatically determined. Moreover, the MOEA is also used to optimize the fuzzy threshold of each node, which enables the proposed method to find diverse overlapping community structures. As a result, the proposed method can adapt to various networks having different overlapping degrees without any predefined parameter.
2) To enhance the performance of the MOEA, a two-stage optimization process is considered in the proposed method. First, the central nodes are optimized by minimizing two objectives (i.e., the kernel $k$-means [26] and the ratio cut [27]), which can find the optimal nonoverlapping communities. Second, the fuzzy thresholds of noncentral nodes are optimized by maximizing two objectives (i.e., the extended modularity [28] and the number of overlapping nodes), which can find a good membership for each noncentral node and, thus, able to detect overlapping communities. Furthermore, two novel population initialization strategies are designed for the two stages, respectively, which are verified to be able to significantly improve the performance of the proposed method.
3) To validate the effectiveness of the proposed method, we implement three approaches by embedding three distance measures in EMOFM, i.e., spectral clustering based Euclidean distance, random-walk-based distance, and diffusion kernel similarity. According to the experimental results, it is empirically verified that the three EMOFM-based approaches outperform six state-of-the-art approaches for overlapping community detection.

The rest of this article is organized as follows. We first give some preliminaries about community detection and related work in Section II. Then, we present the details of the proposed EMOFM in Section III. Afterwards, we report the empirical results to show the effectiveness of the proposed method in Section IV. Finally, Section V concludes this article.

## II. Preliminaries and Related Work

### A. Formal Notations

To start with, the formal notations used in this article are listed in Table I, including those for graphs, fuzzy sets, membership matrix, and MOEA.

TABLE I
Formal Notations Used in This Article

| Notation | Description |
|---|---|
| $G(V,E)$ | a graph $G$ with node set $V$ and edge set $E$ |
| $V$ | a node set $V=\{v_1,v_2,\cdots,v_n\}$ |
| $E$ | a edge set $E=\{(v_i,v_j)\|v_i\in V,v_j\in V,\text{ and } i\neq j\}$ |
| $n,m$ | number of nodes in $V$, number of edges in $E$ |
| $A$ | adjacency matrix of $G$, where $A\in\{0,1\}^{n\times n}$ |
| $d$ | degree of the $i$-th node, where $d_i=\sum_{j=1}^{n}A_{ij}$ |
| $C$ | a community set $C=\{C_1,\cdots,C_k\}$, where $C_1\cup\cdots\cup C_k=V$ |
| $k$ | number of communities in $C$ |
| $\mathbf{b}$ | a binary vector $\mathbf{b}=(b_1,b_2,\cdots,b_n)$, where $\mathbf{b}\in\{0,1\}^n$ |
| $\mathbf{r}$ | a real vector $\mathbf{r}=(r_1,r_2,\cdots,r_n)$, where $\mathbf{r}\in[0,1]^n$ |
| $CN$ | a central node set $CN=\{CN_1,\cdots,CN_k\}$ |
| $NC$ | a non-central node set $NC=\{NC_1,\cdots,NC_{n-k}\}$ |
| $U$ | membership matrix between $NC$ and $CN$, $U\in\mathbb{R}^{(n-k)\times k}$ |
| $U_{ij}$ | membership degree between $NC_i$ and $CN_j$ |
| $P$ | a population $P=\{\mathbf{p}_1,\mathbf{p}_2,\cdots,\}$ |
| $\mathbf{p}_i$ | the $i$-th individual in $P$, $\mathbf{p}_i=\{b_1,\cdots,b_n,r_1,\cdots,r_n\}$ |
| $\mathbf{p}_{i_\mathbf{b}},\mathbf{p}_{i_\mathbf{r}}$ | the vector $(b_1,\cdots,b_n)$ and the vector $(r_1,\cdots,r_n)$ in $\mathbf{p}_i$ |
| $P'$ | a parent population for generating offsprings |
| $O$ | an offspring population |
| $SubP_i$ | the $i$-th subpopulation |
| $N_{pop}$, $N_{sub}$ | size of population $P$, size of each subpopulation $SubP_i$ |
| $Gen$ | maximum number of generations |
| $stage$ | current stage number |
| $FR$ | final result, $FR=P\cup SubP_1\cup\cdots\cup SubP_{\|SubP\|}$ |

### B. Community Detection Problem

Formally, a network can be modeled as a graph denoted by $G=(V,E)$, where $V=\{v_1,v_2,\ldots,v_n\}$ is the set of nodes, $E=\{(v_i,v_j)|v_i\in V,v_j\in V,\text{ and } i\neq j\}$ is the set of edges, and $n=|V|$ is the number of nodes. A network $G$ can also be represented by the adjacency matrix $A=(A_{ij})^{n\times n}$. In this article, we merely consider undirected and unweighted networks, i.e., $A_{ij}=A_{ji}$ and $A_{ij}\in\{0,1\}$. A community structure is defined as $C=\{C_1,C_2,\ldots,C_k\}$ in $G$, which satisfies the following conditions:

$$V=\bigcup_{i=1}^{k}C_i,\ \text{where } C_i\neq\emptyset \tag{1}$$

$$C_i\neq C_j\quad\forall i\neq j\ \text{ and } i,j\in\{1,2,\ldots,k\}. \tag{2}$$

If the set of communities satisfies

$$C_i\cap C_j=\emptyset\quad\forall\, i\neq j\ \text{ and } i,j\in\{1,2,\ldots,k\} \tag{3}$$

then $C$ is a set of nonoverlapping communities of network $G$. By contrast, if the set of communities satisfies

$$C_i\cap C_j\neq\emptyset,\ \exists\, i\neq j\ \text{ and } i,j\in\{1,2,\ldots,k\} \tag{4}$$

then $C$ is a set of overlapping communities.

### C. Fuzzy-Clustering-Based Overlapping Community Detection

Since this article focuses on uncovering overlapping communities, some fuzzy-clustering-based overlapping community detection approaches are reviewed in the following.

In [14], Zhang *et al.* devised an approach based on the fuzzy concept to identify overlapping communities in complex networks. This approach integrates fuzzy clustering with a new modularity function, namely, the spectral mapping, in which an approximation mapping of nodes into Euclidean space is calculated. Then, a soft assignment is obtained by utilizing FCM based on the Euclidean distances between nodes. Finally, the communities can be determined according to a threshold $\lambda$. Note that the number of communities $k$ should be predefined for this approach.

In [17], a distance based on local random walk was applied to overlapping community detection. Specifically, a local distance measure is designed by using the random-walk-based similarity indices. Then, the membership matrix $U$ can be obtained by FCM based on the proposed distance measure. Similar to the approach in [14], the number of communities $k$ and a fuzzy threshold $\lambda$ should be defined in advance to get the final communities.

In contrast to the approaches based on FCM, Zhang *et al.* [29] proposed a nonnegative matrix factorization (NMF) [30] technique for fuzzy community detection, which quantified an absolute membership that a node belongs to a specific community; hence, overlapping communities can be uncovered by using different membership degrees.

Psorakis *et al.* [31] developed a probabilistic approach based on the Bayesian NMF model to detect overlapping communities. In this approach, the feature matrix is decomposed via NMF as part of the parameter inference for a generative model to get a soft assignment. Similarly, an extra fuzzy threshold should be predefined to get the final communities.

Binesh and Rezghi [32] designed a fuzzy clustering algorithm for overlapping community detection based on NMF, which does not need an extra parameter to control the fuzziness in FCM. The experimental results demonstrated that this approach can get better membership matrix than FCM-based approaches with respect to two novel evaluation criteria.

As reported in [11], [22], [23], and [32], most of the existing fuzzy-clustering-based overlapping community detection approaches require some prior knowledge. Specifically, the number of communities $k$ should be given for all the abovementioned approaches, and a fuzzy threshold $\lambda$ should be predefined to get the final communities for the approaches in [14], [17], and [31]. Besides, these approaches define a single fuzzy threshold for all nodes in the network, which is not reasonable since each node has its own trait [23]. By contrast, the proposed EMOFM assigns a separate fuzzy threshold to each node and optimizes both the fuzzy thresholds and the number of communities via an MOEA. The idea of EMOFM is presented in Section III-A.

### D. Evolutionary-Computation-Based Fuzzy Clustering

The fuzzy clustering can be regarded as an optimization problem, where the objective is the quality of the clustering result (e.g., fuzzy $J_m$ index [15], Xie-Beni index [33], and overlap-separation measure [34]), and the variables are the cluster centers. The original FCM optimizes the cluster centers by iteratively updating the membership matrix and the cluster centers. However, this optimization approach is based on hill climbing methods, which are sensitive to the initial cluster centers and likely to get stuck in local optimum [18], [20].

During the last two decades, evolutionary computation has gained popularity in solving various complex optimization problems [35]. It has the following advantages in comparison to the optimization approach in FCM: First, it optimizes a set of individuals (each individual corresponds to a clustering result) simultaneously, which can effectively escape from local optimum. Second, it can optimize not only the cluster centers, but also the parameters (e.g., the number of clusters) in FCM, thus, making FCM parameterless. Third, it can optimize multiple objectives simultaneously (i.e., evolutionary multiobjective optimization), which can obtain better clustering result than optimizing one single objective [36]. As a result, a number of evolutionary algorithms have been adopted in fuzzy clustering, such as genetic algorithm [18], [37], teaching-learning-based optimization [38], particle swarm optimization [19], [20], [36], [39].

In [19], Wang *et al.* suggested a particle swarm optimization-based FCM for overlapping community detection. Specifically, it embeds FCM in the selection operator of particle swarm optimization and uses it to optimize the cluster centers by minimizing the fuzzy $J_m$ index. However, the number of clusters should be given in advance.

Saha *et al.* [36] proposed a multiobjective differential evolution-based fuzzy clustering technique, which optimizes the cluster centers by minimizing both the fuzzy $J_m$ index and the Xie-Beni index. However, the number of cluster centers should also be predefined.

To automatically determine the number of cluster centers, a multiobjective genetic algorithm NSGA-II was incorporated into FCM by Wikaisuksakul [18]. This approach also optimizes the cluster centers, but each individual has a different length so that it can represent different numbers of cluster centers. As for the objectives, it minimizes both the fuzzy $J_m$ index and the overlap-separation measure.

In contrast to the abovementioned approaches that optimize the cluster centers, a multiobjective teaching-learning-based optimization algorithm was introduced in [38] to optimize the membership matrix. This approach minimizes the fuzzy $J_m$ index and maximizes the partition coefficient and exponential separation, as the interaction between these two objectives is important to dynamically determine the number of clusters and explore interesting areas of the search space.

Although various evolutionary algorithms have been applied to fuzzy clustering, most of them cannot be used for overlapping community detection due to the following three reasons. First, the nodes in complex networks are not in Euclidean space; hence, it is difficult to represent the cluster centers. Second, the fuzzy threshold of each node is needed to obtain the overlapping communities, which should be optimized together with the cluster centers. Third, existing evolutionary algorithms are ineffective for fuzzy-clustering-based overlapping community detection due to the difficulty of the problem. To address these issues, we adopt the idea of fuzzy $c$-medoids that forces each cluster center to be an existing node and optimize both the cluster centers and the fuzzy thresholds. We also enhance the performance of an existing evolutionary algorithm by a two-stage optimization process and two-population initialization strategies. The procedure of the proposed EMOFM is elaborated in Section III-B.

## III. Proposed Fuzzy Method

### A. Main Idea of EMOFM

In the proposed EMOFM, each cluster center is an existing node in the network $G(V, E)$ to be clustered, termed central node. To begin with, a binary vector $\mathbf{b}$ and a real vector $\mathbf{r}$ are

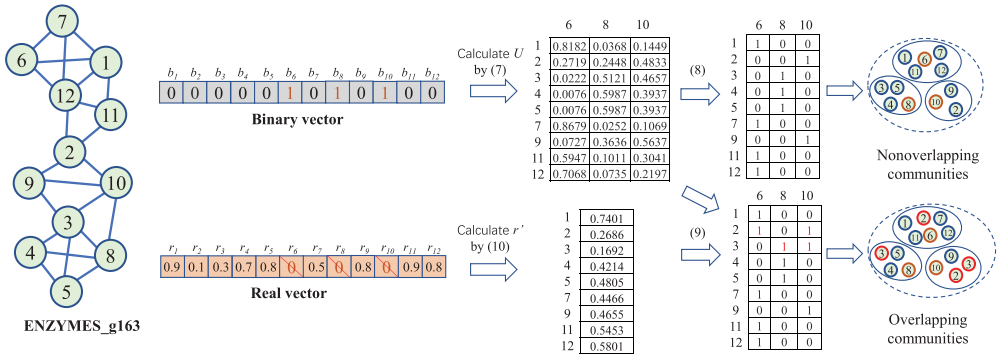


Fig. 1. Illustrative example of the proposed method to detect the communities of the ENZYMES_g163 network. The values of the binary vector and real vector are randomly generated for illustrative purposes, while they are related to the first and second optimization stages of EMOFM, respectively.

defined as follows:

$$\begin{aligned}\mathbf{b} &= (b_1, b_2, \ldots, b_n) \in \{0,1\}^n \\ \mathbf{r} &= (r_1, r_2, \ldots, r_n) \in [0,1]^n \end{aligned} \tag{5}$$

where $b_i$ indicates whether $v_i$ is a central node or not, $r_i$ denotes the fuzzy threshold of $v_i$, and $n$ denotes the number of nodes of the network.

To obtain the communities, the noncentral node set NC and the central node set CN are first identified, i.e.,

$$\begin{aligned}\text{NC} &= \{\text{NC}_1, \text{NC}_2, \cdots\} \\ \text{CN} &= \{\text{CN}_1, \text{CN}_2, \cdots\} \end{aligned} \tag{6}$$

where $\text{NC}_i$ is the $i$th node whose corresponding element in $\mathbf{b}$ is 0, and $\text{CN}_i$ is the $i$th node whose corresponding element in $\mathbf{b}$ is 1. Then, the distances between noncentral nodes and central nodes are calculated based on a distance measure, and each element in the membership matrix $U = (U_{ij})^{|\text{NC}|\times|\text{CN}|}$ can be obtained by [15]

$$U_{ij} = \frac{1}{\sum_{l=1}^{|\text{CN}|} \frac{\text{dis}(\text{NC}_i, \text{CN}_j)}{\text{dis}(\text{NC}_i, \text{CN}_l)}^{\frac{2}{f-1}}} \tag{7}$$

where $U_{ij}$ denotes the membership degree between $\text{NC}_i$ and $\text{CN}_j$, $\text{dis}(\text{NC}_i, \text{CN}_j)$ denotes the distance between the two nodes, and $f$ is a parameter controlling the fuzziness of the resulting clusters, which was suggested to be 2 in [32]. Last, in order to obtain the nonoverlapping communities, each central node is regarded as the center of a community, and each noncentral node $\text{NC}_i$ belongs to the community centered at the central node $\text{CN}_j$ if

$$U_{ij} = \max_{l=1,\ldots,|\text{CN}|} U_{il}. \tag{8}$$

On the other hand, to obtain the overlapping communities, each noncentral node $\text{NC}_i$ belongs to the community centered at the central node $\text{CN}_j$ once

$$U_{ij} \geq r'_{\text{NC}_i} \tag{9}$$

where $r'_{\text{NC}_i}$ is obtained by

$$r'_{\text{NC}_i} = \min_l U_{il} + r_{\text{NC}_i} \times \left( \max_l U_{il} - \min_l U_{il} \right). \tag{10}$$

As a consequence, the proposed method can detect overlapping nodes by finding proper fuzzy thresholds, where a lower value of $r_{\text{NC}_i}$ indicates a higher probability that $\text{NC}_i$ is an overlapping node, and vice versa. In extreme cases, $r_{\text{NC}_i} = 1$ means that $\text{NC}_i$ belongs to a single community, and $r_{\text{NC}_i} = 0$ means that $\text{NC}_i$ belongs to all communities.

To better understand the abovementioned procedure, Fig. 1 gives an illustrative example of the proposed method to detect the communities of the cheminformatics network ENZYMES_g163 [40], which has 12 nodes and 22 edges. Assume that the binary vector and the real vector are $\mathbf{b} = (b_1, \ldots, b_{12}) = (0,0,0,0,0,1,0,1,0,1,0,0)$ and $\mathbf{r} = (r_1, \ldots, r_{12}) = (0.9, 0.1, 0.3, 0.7, 0.8, 0, 0.5, 0, 0.8, 0, 0.9, 0.8)$, respectively, where the values of $\mathbf{b}$ and $\mathbf{r}$ are randomly generated for illustrative purposes. The membership matrix $U$ can be calculated by (7), and $r'$ can be calculated by (10). Afterwards, the nonoverlapping communities and the overlapping communities can be obtained by (8) and (9), respectively. As shown in Fig. 1, the nonoverlapping communities of the network are $\{\{3,4,5,8\},\{1,6,7,11,12\},\{2,9,10\}\}$, while the overlapping communities are $\{\{3,4,5,8\},\{1,2,6,7,11,12\},\{2,3,9,10\}\}$ with nodes 2 and 3 being overlapping nodes. To sum up, the idea of the proposed method is entirely based on fuzzy clustering. On one hand, it calculates the membership matrix $U$ between noncentral nodes and central nodes, where $U$ represents the membership degrees between nodes and cluster centers. On the other hand, it assigns a fuzzy threshold to each node, thus, controlling whether the node is an overlapping node or not.

In the following, three methods adopted by EMOFM to measure the distances between nodes are introduced.

*1) Spectral-clustering-based Euclidean distance:* Spectral clustering algorithm is a clustering method based on globally optimizing cost functions [41]. In [42], the spectral clustering algorithm was first adopted in the detection of nonoverlapping communities. Later, it was adopted to detect overlapping communities by replacing $k$-means in spectral clustering with FCM [14]. To calculate the spectral-clustering-based Euclidean distance, the adjacent matrix $A$ and the degree matrix $D$ of the network should be obtained, where $D$ is a diagonal matrix with each element of the principal diagonal being the degree of a node. Then, the Laplace matrix can be

constructed by

$$L = D - A \tag{11}$$

and the eigenvectors corresponding to the top $k-1$ minimum eigenvalues of $L$ are calculated. Afterwards, a new matrix $\text{Eig} \in \mathbb{R}^{n\times(k-1)}$ is obtained by combining all the $k-1$ eigenvectors. Finally, each row of Eig is regarded as the feature of a node, and the distance between two nodes can be obtained by calculating the Euclidean distance between the corresponding features.

*2) Random-walk-based distance:* Random walk is a stochastic process that starts from an initial node and goes to adjacent nodes with a probability [43]. Based on Markov chains, this process can be modeled by the average commute time (ACT) [44], which defines the time of a random walk starting from a node, arriving at another node for the first time, and going back to the node. As reported in [45] and [46], the ACT can be used to measure the distance between any two nodes in a network. The ACT-based distance between two nodes $\text{ACT}(v_i, v_j)$ can be calculated by [44]

$$\text{ACT}(v_i, v_j) = 2m(\mathbf{u}_i - \mathbf{u}_j)L^+(\mathbf{u}_i - \mathbf{u}_j)^T \tag{12}$$

where $\mathbf{u}_i \in \mathbb{R}^{1\times n}$ is a vector of zeros with only the $i$th element being 1, and $L^+$ is the Moore–Penrose pseudoinverse of the Laplace matrix $L$, which can be obtained by [47]

$$L^+ = \left(L - \frac{II^T}{n}\right)^{-1} + \frac{II^T}{n} \tag{13}$$

where $I \in \mathbb{R}^{n\times 1}$ is a vector of ones.

*3) Diffusion kernel similarity:* Since the distance between two nodes is usually negatively related to the similarity between them, the similarity can be used to measure the distance between nodes in a network. In [48], the diffusion kernels on networks were tailored for kernel-based learning algorithms. Diffusion kernels similarity is a kind of exponential kernels based on heat equation, which can be calculated by [48]

$$K = e^{-\beta\times L} \tag{14}$$

where $\beta$ is suggested to be 1 and $L$ is the Laplace matrix. Note that since the diffusion kernels similarity is negatively related to the distance, we use the dissimilarity $DK = (DK_{ij})^{n\times n}$ to measure the distance between nodes, i.e.,

$$DK = \max(K) - K \tag{15}$$

where $\max(K)$ denotes the maximum value in $K$, and the element $DK_{ij}$ denotes the distance between nodes $v_i$ and $v_j$.

### *B. Optimization Process of EMOFM*

As shown in Fig. 2, the proposed EMOFM contains two stages: The first stage aims to uncover the nonoverlapping communities by finding the optimal central nodes of communities, while the second stage detects the overlapping nodes by finding the optimal fuzzy thresholds of noncentral nodes. Both two stages are implemented based on an effective MOEA called AR-MOEA [24], but the objectives to be optimized in the two stages are different. Specifically, the first stage optimizes two objectives related to nonoverlapping communities, and the second stage optimizes two objectives related to overlapping communities.

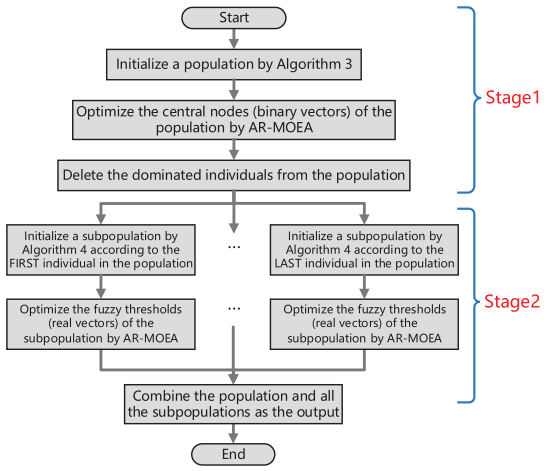


Fig. 2. Optimization process of EMOFM.

**Algorithm 1:** General Procedure of EMOFM.

**Input:** $Gen$: Maximum number of generations for each stage; $N_{pop}$: Population size; $N_{sub}$: Subpopulation size;
**Output:** $FR$: Final result;
**Stage 1: Uncovering nonoverlapping communities**
1: $P \leftarrow Initialization(N_{pop})$; // Algorithm 3
2: $P \leftarrow AR\_MOEA(P, Gen, 1)$; // Algorithm 2
3: Delete all dominated individuals from $P$;
**Stage 2: Detecting overlapping nodes**
4: **for** $i = 1$ to $|P|$ **do**
5: $SubP_i \leftarrow SubInitialization(\mathbf{p}_i, N_{sub})$; // Algorithm 4
6: $SubP_i \leftarrow AR\_MOEA(SubP_i, Gen, 2)$; // Algorithm 2
7: Delete all dominated individuals from $SubP_i$;
8: **end for**
9: $FR \leftarrow SubP_1 \cup \cdots \cup SubP_{|P|}$;
10: **return** $FR$;

The procedure of the proposed EMOFM is summarized in Algorithm 1. In the first stage, EMOFM generates an initial population $P$ by an initialization strategy (see Line 1), where each individual $\mathbf{p}_i$ in the population consists of a binary vector $\mathbf{p}_{i_\text{b}}$ and a real vector $\mathbf{p}_{i_\text{r}}$. Then, it optimizes the center nodes (represented by the binary vector) according to the following two widely used objectives for Gen generations (see Line 2):

$$\min \begin{cases} \text{KKM} = 2(n-k) - \sum_{i=1}^{k} \dfrac{L(C_i, C_i)}{|C_i|} \\ \text{RC} = \sum_{i=1}^{k} \dfrac{L(C_i, V\backslash C_i)}{|C_i|} \end{cases} \tag{16}$$

where $n$ is the number of nodes of network $G(V, E)$, $k$ is the number of communities, $C_i$ denotes the $i$th community,

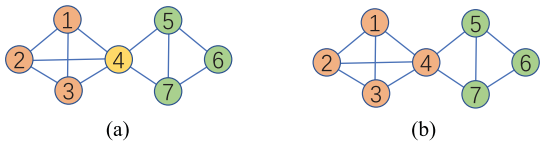


Fig. 3. Two community structures of a network, where the $Q_{ov}$ value will increase if node 4 is regarded as a nonoverlapping node. (a) Overlapping communities with $Q_{ov} = 0.1667$. (b) Nonoverlapping communities with $Q_{ov} = 0.1875$.

$L(C_i, C_j) = \sum_{v \in C_i, w \in C_j} A_{vw}$, and $A$ is the adjacent matrix of the network. The first objective KKM (i.e., kernel $k$-means [26]) is the intralink density in all communities, and the second objective RC (i.e., ratio cut [27]) is the interlink density between different communities; hence, it is expected to obtain a community structure with high intralink density and low interlink density by optimizing these two objectives. Note that the first stage aims to uncover the nonoverlapping communities, which are obtained by (8) according to each individual.

In the second stage, EMOFM first generates one subpopulation Sub$P_i$ based on the $i$th nondominated individual in $P$ (see Line 5). Afterwards, for each subpopulation Sub$P_i$, EMOFM optimizes the fuzzy thresholds (represented by the real vector) according to the following two objectives for Gen generations (see Line 6):

$$\max \begin{cases} f_1 = Q_{ov} \\ f_2 = \text{Num}_{\text{overlapping}} \end{cases} \tag{17}$$

where $\text{Num}_{\text{overlapping}}$ is the number of overlapping nodes, and $Q_{ov}$ is the extended modularity [28] calculated as

$$Q_{ov} = \frac{1}{2m} \sum_{i=1}^{k} \sum_{v \in C_i, w \in C_i} \frac{1}{O_v O_w} \left[ A_{vw} - \frac{d_v d_w}{2m} \right] \tag{18}$$

where $m$ is the number of edges in the network, $O_v$ is the number of communities which node $v$ belongs to, and $d_v$ is the degree of node $v$. The larger the value of $Q_{ov}$, the better the quality of detected overlapping communities; hence, it can be used to guide EMOFM to detect the overlapping nodes. The reason for maximizing the number of overlapping nodes is that not all meaningful overlapping nodes can be found by maximizing only $Q_{ov}$. Let us consider the network shown in Fig. 3, where node 4 is likely to be an overlapping node. However, a larger $Q_{ov}$ value can be obtained if node 4 is regarded as a nonoverlapping node. Note that the overlapping communities are obtained by (9) for the calculation of $Q_{ov}$.

The optimization process of AR-MOEA is detailed in Algorithm 2. Different from the original AR-MOEA, the AR-MOEA adopted here optimizes different objective functions in the two stages. More specifically, it evaluates individuals by (16) in the first stage and by (17) in the second stage (see Lines 1–5), and it evolves the binary vectors in the first stage and real vectors in the second stage (see Lines 12–20). Besides, the mating selection (see Line 7) is to select a set of parents with good convergence and diversity from $P$ for generating offsprings, and the environmental selection (see Line 23) is to retain some individuals with good convergence and diversity from $P \cup O$ for the next generation. Both the mating selection and environmental selection are based on an enhanced inverted generational distance-based selection with adaptive reference points, the details of which are elaborated in [24].

**Algorithm 2:** $AR\_MOEA(P, Gen, stage)$.

**Input:** $P$: Population to be optimized; $Gen$: Maximum number of generations; $stage$: Current stage number;
**Output:** $P$: Optimized population;

1: **if** $stage == 1$ **then**
2: Evaluate the individuals in $P$ by (16);
3: **else if** $stage == 2$ **then**
4: Evaluate the individuals in $P$ by (17);
5: **end if**
6: **for** $g = 1$ to $Gen$ **do**
7: $P' \leftarrow MatingSelection(P)$;
8: $O \leftarrow \emptyset$;
9: **while** $P' \neq \emptyset$ **do**
10: Randomly select two individuals $\mathbf{p}_i$ and $\mathbf{p}_j$ from $P'$;
11: $P' \leftarrow P' \setminus \{\mathbf{p}_i, \mathbf{p}_j\}$;
12: **if** $stage == 1$ **then**
13: $[\mathbf{b}_1, \mathbf{b}_2] \leftarrow$ Perform binary genetic operators on $\mathbf{p}_{i_\mathbf{b}}$ and $\mathbf{p}_{j_\mathbf{b}}$;
14: $\mathbf{p}_{i_\mathbf{b}} \leftarrow \mathbf{b}_1$, $\mathbf{p}_{j_\mathbf{b}} \leftarrow \mathbf{b}_2$;
15: Evaluate the modified $\mathbf{p}_i$ and $\mathbf{p}_j$ by (16);
16: **else if** $stage == 2$ **then**
17: $[\mathbf{r}_1, \mathbf{r}_2] \leftarrow$ Perform real genetic operators on $\mathbf{p}_{i_\mathbf{r}}$ and $\mathbf{p}_{j_\mathbf{r}}$;
18: $\mathbf{p}_{i_\mathbf{r}} \leftarrow \mathbf{r}_1$, $\mathbf{p}_{j_\mathbf{r}} \leftarrow \mathbf{r}_2$;
19: Evaluate the modified $\mathbf{p}_i$ and $\mathbf{p}_j$ by (17);
20: **end if**
21: $O \leftarrow O \cup \{\mathbf{p}_i, \mathbf{p}_j\}$;
22: **end while**
23: $P \leftarrow EnvironmentalSelection(P \cup O)$;
24: **end for**
25: **return** $P$;

Finally, as shown in Line 9 of Algorithm 1, EMOFM returns the population obtained in the first stage and all the subpopulations obtained in the second stage as the final result. In the next two sections, the population initialization strategies in the two stages of EMOFM are elaborated, respectively.

### C. Population Initialization Strategy in the First Stage

The population initialization strategy in the first stage aims to generate a set of initial individuals, such that some central nodes in these individuals have a large probability of being the centers of real communities. For this aim, we consider the nodes with larger degrees as candidate central nodes, since the nodes in the center of a community are likely to have more connections to the others. Therefore, the nodes in the initial individuals with larger degrees are more likely to be central nodes (i.e., the corresponding binary variables are set to 1).

**Algorithm 3:** $Initialization(N_{pop})$.

**Input:** $N_{pop}$: Population size;
**Output:** $P$: Initial population;
**Step 1: Find candidate central nodes**
1: $V \leftarrow \{1, 2, \ldots, n\}$; // Set of remaining nodes
2: $K \leftarrow \emptyset$; // Set of candidate central nodes
3: **while** $V \neq \emptyset$ **do**
4: $w \leftarrow \text{argmax}_{v \in V} d_v$; // $d_v$ is the degree of node $v$
5: $K \leftarrow K \cup \{w\}$;
6: Delete $w$ and all its adjacent nodes from $V$;
7: **end while**
**Step 2: Initialize the population**
8: $P \leftarrow N_{pop}$ individuals with all variables being 0;
9: **for** $i = 1$ to $N_{pop}/2$ **do**
10: $rand \leftarrow$ A random integer within $[1, |K|]$;
11: Randomly select $rand$ nodes from $K$ and set the corresponding binary variables of $\mathbf{p}_{i_\mathbf{b}}$ to 1;
12: **end for**
13: **for** $i = N_{pop}/2 + 1$ to $N_{pop}$ **do**
14: $rand \leftarrow$ A random integer within $[1, n]$;
15: Randomly select $rand$ nodes from $\{1, 2, \ldots, n\}$ and set the corresponding binary variables of $\mathbf{p}_{i_\mathbf{b}}$ to 1;
16: **end for**
17: **return** $P$;

The detailed procedure of the population initialization strategy is shown in Algorithm 3. To find the candidate central nodes, the node with the maximum degree is selected as a candidate central node, and all its adjacent nodes are ignored. After that, the node with the maximum degree among the remaining nodes is selected as a candidate central node, and this procedure repeats until all the nodes are selected or ignored (see Lines 3–7).

Afterwards, the initial population $P$ can be generated based on the candidate central nodes, where some candidate central nodes are randomly selected for each individual in $P$, and the values of corresponding binary variables of the individual are set to 1 (see Lines 9–12). Note that in order to improve the diversity of the initial population as well as alleviate premature convergence, only half the individuals in $P$ are generated by the abovementioned strategy, while the other half are generated by randomly selecting nodes in the network as central nodes (see Lines 13–16).

### *D. Subpopulation Initialization Strategy in the Second Stage*

After optimizing the center nodes in the first stage, the fuzzy thresholds will be optimized in the second stage. However, due to the difference between the center nodes in the optimized individuals, they cannot be optimized together in the second stage. In other words, the fuzzy thresholds of noncentral nodes for each individual in $P$ should be optimized separately. To this end, a subpopulation Sub$P_i$ needs to be generated for each individual $\mathbf{p}_i$, so that the fuzzy thresholds of noncentral nodes for $\mathbf{p}_i$ can be optimized by evolving Sub$P_i$.

The detailed procedure of the subpopulation initialization strategy in the second stage is given in Algorithm 4, which

**Algorithm 4:** $Subinitialization(\mathbf{p}_i, N_{sub})$.

**Input:** $\mathbf{p}_i$: The $i$-th individual; $N_{sub}$: Subpopulation size;
**Output:** $SubP_i$: The $i$-th subpopulation;
**Step1: Initialize fuzzy thresholds**
1: $[NC, CN] \leftarrow$ Determine the set of noncentral nodes and the set of central nodes by (6) according to $\mathbf{p}_{i_\mathbf{b}}$;
2: $U \leftarrow$ Calculate the membership matrix between $NC$ and $CN$ by (7);
3: $[r_1, r_2, \ldots, r_n] \leftarrow 0$; // Initial fuzzy thresholds
4: **for** $j = 1$ to $|NC|$ **do**
5: $[S_1, S_2] \leftarrow$ Use $k$-means to cluster $U_{j1}, U_{j2}, \ldots, U_{j|CN|}$ into two sets;
6: $r_{NC_j} \leftarrow$ Calculate the fuzzy threshold of $NC_j$ by (19);
7: **end for**
**Step2: Initialize subpopulations**
8: $SubP_i \leftarrow \emptyset$;
9: **for** $t = 1$ to $N_{sub}$ **do**
10: $\mathbf{p}_{i_\mathbf{r}} \leftarrow [r_1, r_2, \ldots, r_n]$;
11: **for** $j = 1$ to $n$ **do**
12: $rand \leftarrow$ A random value within [0,1];
13: **if** $rand < 0.5$ **then**
14: Set the $j$-th real variable of $\mathbf{p}_{i_\mathbf{r}}$ to a random value within [0,1];
15: **end if**
16: **end for**
17: $SubP_i \leftarrow SubP_i \cup \{\mathbf{p}_i\}$;
18: **end for**
19: **return** $SubP_i$;

consists of two steps, i.e., initialization of fuzzy thresholds and initialization of subpopulations. To initialize the fuzzy thresholds for individual $\mathbf{p}_i$, the membership matrix $U$ between NC and CN is first calculated according to the binary vector $\mathbf{p}_{i_\mathbf{b}}$ (see Line 2). Then, for each noncentral node NC$_j$, the membership degrees between NC$_j$ and all central nodes $U_{j1}, U_{j2}, \ldots, U_{j|\text{CN}|}$ are clustered into two groups $S_1$ and $S_2$, where the values in $S_1$ are smaller than those in $S_2$ (see Line 5). Finally, the minimum value in $S_2$ is regarded as $r'_{\text{NC}_j}$, and $r_{\text{NC}_j}$ can be calculated by the inverse operation of (10) (see Line 6), i.e.,

$$r_{\text{NC}_j} = \frac{r'_{\text{NC}_j} - \min_l U_{jl}}{\max_l U_{jl} - \min_l U_{jl}}. \tag{19}$$

An example of initializing fuzzy thresholds on the ENZYMES_g163 network is given in Fig. 4, where the values of $U$ are randomly generated for illustrative purposes.

After initializing the fuzzy thresholds of noncentral nodes for individual $\mathbf{p}_i$, the subpopulation Sub$P_i$ can be generated, where the real vectors of all the individuals in Sub$P_i$ are the same to the fuzzy thresholds $r_1, r_2, \ldots, r_n$ (see Line 10). To increase the diversity of Sub$P_i$, each real variable of the individuals in Sub$P_i$ is randomly perturbed with a probability of 0.5 (see Lines 12–15). As a consequence, the individuals in the subpopulation Sub$P_i$ have the same central nodes as individual $\mathbf{p}_i$, so these individuals can be evolved together. On the other hand, these

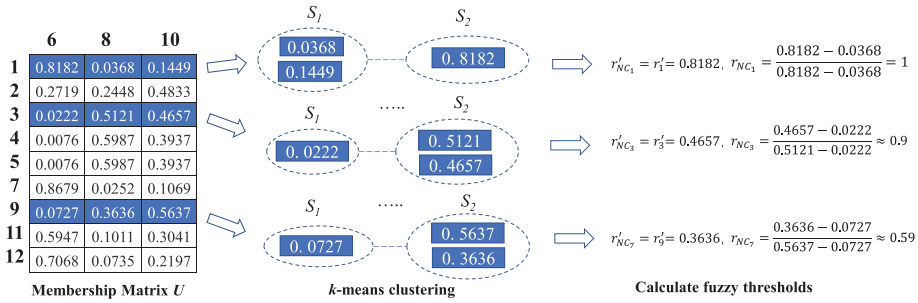


Fig. 4. Illustrative example of the initialization of fuzzy thresholds on the ENZYMES_g163 network. The values of the membership matrix are randomly generated for illustrative purposes.

individuals have diverse real vectors that are generated by perturbing the initial fuzzy thresholds, and better fuzzy thresholds for the noncentral nodes are expected to be obtained by evolving these individuals.

### E. Complexity Analysis

The time complexity of the proposed EMOFM is mainly determined by two components, i.e., the calculation of objective functions and the optimization process of AR-MOEA. According to the definitions of KKM, RC, and $Q_{ov}$, the time complexities for calculating them are $O(n+m)$, $O(n+m)$, and $O(n^2)$, respectively, where $n$ is the number of nodes and $m$ is the number of edges of the network. The time complexity of one generation of AR-MOEA is $O(N^3)$ [24], where $N$ is the population size. Therefore, the worst time complexity of EMOFM in the first stage is $O((n+m)N_{\text{pop}}\text{Gen})+O(N_{\text{pop}}^3\text{Gen})$ and the worst time complexity of EMOFM in the second stage is $O(n^2N_{\text{sub}}\text{Gen}N_{\text{pop}})+O(N_{\text{sub}}^3\text{Gen}N_{\text{pop}})$, where $N_{\text{pop}}$ is the population size in the first stage, $N_{\text{sub}}$ is the subpopulation size in the second stage, and Gen is the maximum number of generations for each stage. To summarize, the total time complexity of EMOFM is $O((n+m+N_{\text{pop}}^2+n^2N_{\text{sub}}+N_{\text{sub}}^3)N_{\text{pop}}\text{Gen})$. Since $n \gg N_{\text{pop}}$, $n \gg N_{\text{sub}}$ and $n^2 \gg m$ in general, the time complexity of EMOFM can be considered as $O(n^2N_{\text{sub}}N_{\text{pop}}\text{Gen})$.

On the other hand, the space complexity of the proposed EMOFM is mainly determined by three variables, i.e., the distance matrix between each two nodes, the population in the first stage, and the subpopulations in the second stage, the space complexities of which are $O(n^2)$, $O(nN_{\text{pop}})$, and $O(nN_{\text{sub}}N_{\text{pop}})$, respectively. Therefore, the space complexity of EMOFM is $O(n(n+N_{\text{sub}}N_{\text{pop}}))$.

## IV. Empirical Results

In this section, the performance of the proposed EMOFM is verified by comparing it with six state-of-the-art approaches for overlapping community detection, namely, the local maximum degree nodes based approach (LMD) [8], Zhang's approach [49], the NMF approach [50], the improved multiobjective quantum-behaved particle swarm optimization (IMOQPSO) [51], the maximal clique-based multiobjective evolutionary algorithm (MCMOEA) [52], and the mixed representation based multiobjective evolutionary algorithm (MR-MOEA) [12]. Among the six baseline approaches, LMD is based on local expansion, Zhang's approach is based on core nodes, NMF is based on fuzzy clustering, and IMOQPSO, MCMOEA, and MR-MOEA are based on evolutionary multiobjective optimization. In addition, we embed the three distance measures introduced in Section III-A in EMOFM, which are hereafter denoted as EMOFM-SC (with spectral-clustering-based Euclidean distance), EMOFM-RW (with random-walk-based distance), and EMOFM-DK (with diffusion kernel similarity), respectively.

In the rest of this section, the performance of EMOFM on real-world networks and LFR benchmark networks are verified in Section IV-B and Section IV-C, respectively, the effectiveness of the population initialization strategy in EMOFM is verified in Section IV-D, and the effectiveness of the two-stage optimization process in EMOFM is verified in Section IV-E. Last, the limitations of EMOFM are discussed in Section IV-F.

### A. Experimental Settings

*1) Parameter settings of approaches:* The threshold controlling the merge of communities in LMD is set to 0.4, and the parameters in the other baseline approaches are set as suggested in their original literature. For EMOFM, the population size $N_{\text{pop}}$ is set to 100, the subpopulation size $N_{\text{sub}}$ is set to 10, the maximum number of generations Gen is set to 100. The uniform crossover and bitwise mutation are adopted in the first stage, and the simulated binary crossover [53] and polynomial mutation [54] are adopted in the second stage. The probability of crossover is set to 1, the probability of mutation is set to $1/n$ ($n$ denotes the number of nodes in the network), and the distribution index of crossover and mutation is set to 20. As for the other MOEA-based approaches, the population size and the number of function evaluations are set to the same to those in EMOFM. Besides, Table II lists the representations, genetic operators, and relevant parameter settings in all the MOEA-based approaches.

*2) Real-world networks:* As listed in Table III, 16 well-known and widely used real-world networks are adopted in the experiments. These networks are extracted from various domains and with different scales, degree distributions, and characteristics, which can not only assess the performance of community detection approaches, but also challenge them in terms of robustness and scalability.

*3) LFR benchmark networks:* A number of synthetic networks produced by the Lancichinetti–Fortunato–Radicchi

TABLE II
REPRESENTATIONS AND OPERATORS USED IN THE MOEA-BASED APPROACHES

| Algorithm | Representation | Operators | Parameter settings |
|---|---|---|---|
| IMOQPSO | Line graph based representation | Improved real based quantum-behaved particle swarm optimization | Harmony memory considering rate is set to a random value in $[0.5, 0.9]$, pitch adjusting rate is set to a random value in $[0.3, 0.5]$, bandwidth is set to a random value in $[0.05, 0.1]$, and the number of measurements is set to 5 |
| MC-MOEA | Maximal clique based representation | One-way crossover and improved label based mutation | Crossover probability is set to 0.7 |
| MR-MOEA | Overlapping node based representation | Improved integer based particle swarm optimization | Inertia weight is set to a random value in [0,1] and both cognitive component and social component are set to 1.494 |
| EMOFM | Central node and fuzzy threshold based representation | Uniform crossover, bitwise mutation, simulated binary crossover, and polynomial mutation | Crossover probability is set to 1, mutation probability is set to $1/n$, and distribution index of both crossover and mutation is set to 20 ($n$ denotes the number of nodes) |

TABLE III
CHARACTERISTICS OF 16 REAL-WORLD NETWORKS

| Network | Nodes | Edges | Avg. degree | Communities | Type |
|---|---|---|---|---|---|
| ENZYMES_g163 [40][1] | 12 | 22 | 3.67 | unknown | Cheminformatics |
| Karate [55][2] | 34 | 78 | 4.69 | 2 | Social |
| Dolphin [56][2] | 62 | 159 | 5.13 | 2 | Social |
| Polbook [57][2] | 105 | 441 | 8.4 | 3 | Social |
| Football [57][2] | 115 | 613 | 10.66 | 12 | Social |
| Email [21][3] | 1133 | 5451 | 4.81 | unknown | Social |
| Blogs [12][2] | 3984 | 6803 | 3.41 | unknown | Social |
| ego-Facebook [32][6] | 4039 | 88234 | 21.85 | unknown | Social |
| SFI [27][4] | 118 | 200 | 1.69 | unknown | Collaboration |
| Jazz [58][3] | 198 | 2742 | 27.70 | unknown | Collaboration |
| Netscience [27][2] | 1589 | 2742 | 3.45 | unknown | Collaboration |
| Erdös [59][7] | 6927 | 11850 | 1.71 | unknown | Collaboration |
| Celegans metabolic [59][3] | 453 | 2040 | 4.50 | unknown | Biology |
| Yeast-D2 [60][5] | 1443 | 6993 | 9.69 | 162 | Biology |
| Y2H [61][5] | 1966 | 2705 | 2.75 | 203 | Biology |
| PPI [9] | 2445 | 6265 | 5.12 | unknown | Biology |

[1][Online]. Available: http://networkrepository.com/chem.php
[2][Online]. Available: http://www-personal.umich.edu/ mejn/netdata/
[3][Online]. Available: http://deim.urv.cat/ alexandre.arenas/data/welcome.htm
[4][Online]. Available: http://dsec.pku.edu.cn/ jliu/
[5][Online]. Available: http://faculty.uaeu.ac.ae/nzaki/ProRank.htm
[6][Online]. Available: http://snap.stanford.edu/data/
[7][Online]. Available: http://vlado.fmf.uni-lj.si/pub/networks/pajek/data/gphs.htm

TABLE IV
PARAMETER SETTINGS OF LFR NETWORKS

| $n$ | $\mu$ | $O_n$ | $O_m$ | $c_{min}$ | $c_{max}$ | Others |
|---|---|---|---|---|---|---|
| 100 | {0.1,0.2,0.3,0.4,0.5,0.6} | $0.1n$ | {2,4,6,8} | 5 | 12 | $k$=10 |
| 500 | {0.1,0.2,0.3,0.4,0.5,0.6} | $0.1n$ | {2,4,6,8} | 10 | 50 | $k_{max}$=50 |
| 1000 | {0.1,0.2,0.3,0.4,0.5,0.6} | $0.1n$ | {2,4,6,8} | 10 | 50 | $\tau_1$=2, $\tau_2$=1 |

(LFR) model [62] are also adopted to evaluate the performance of the compared approaches. The LFR networks follow the settings in [52], which are listed in Table IV. Specifically, the number of nodes $n$ is set to 100, 500, and 1000, the mixing parameter $\mu$ varies from 0.1 to 0.6, the overlapping membership $O_m$ varies from 2 to 8, the number of overlapping nodes $O_n$ is set to $0.1n$, the minimum number of nodes in each community is set to 5 and 10, the maximum number of nodes in each community is set to 12 and 50, and the other parameters $k$, $k_{\max}$, $\tau_1$, and $\tau_2$ are fixed to 10, 50, 2, and 1, respectively. As a result, there are 72 LFR networks with different scales and characteristics.

*4) Performance metrics:* Two popular metrics are used to quantitatively compare the quality of overlapping communities obtained by different approaches on the benchmark networks, i.e., the extended modularity ($Q_{ov}$) [28] and the generalized normalized mutual information (gNMI) [63]. The definition of $Q_{ov}$ can be found in (18), where a larger value of $Q_{ov}$ indicates a better overlapping community structure. On the other hand, gNMI is an information-theoretic measure of the agreement between two community structures, which is used to evaluate the agreement between the obtained communities and the ground truth. The gNMI between a ground truth CA and the communities obtained by an approach CB can be mathematically defined as

$$\text{gNMI}(\text{CA},\text{CB}) = \frac{-2\sum_{i=1}^{k_{\text{CA}}}\sum_{j=1}^{k_{\text{CB}}} N_{ij}\log(\frac{nN_{ij}}{N_{i.}N_{.j}})}{\sum_{i=1}^{k_{\text{CA}}} N_{i.}\log(\frac{N_{i.}}{n}) + \sum_{j=1}^{k_{\text{CB}}} N_{.j}\log(\frac{N_{.j}}{n})} \tag{20}$$

where $k_{\text{CA}}$ denotes the number of communities in CA, $k_{\text{CB}}$ denotes the number of communities in CB, $N \in \mathbb{R}^{k_{\text{CA}}\times k_{\text{CB}}}$ is a matrix with $N_{ij}$ being the number of nodes in the $i$th community of CA that appear in the $j$th community of CB, $N_{i.}$ is the sum over the $i$th row of $N$, and $N_{.j}$ is the sum over the $j$th column of $N$. A larger value of gNMI indicates a higher similarity between CA and CB. It is worth noting that the evolutionary multiobjective optimization-based approaches can obtain multiple overlapping community structures in a single run, we select the one having the best metric value as the final output as suggested in [27], [51], and [52].

Each approach is executed for 30 independent runs on each network, and the best value, average value, and standard deviation of the metric values are recorded. All the experiments are carried out on a computer with Intel I5-6400 2.70 GHz CPU, 8 GB RAM, and Windows 10 operating system.

### *B. Results on Real-World Networks*

The $Q_{ov}$ values obtained by six baseline approaches and three EMOFM-based approaches on 16 real-world networks are listed in Table V. Note that the results of Zhang's approach, IMOQPSO, and MCMOEA on some networks cannot be obtained due to the limitation of internal storage. As can be seen, the fuzzy-clustering-based approach NMF and the MOEA-based approach MR-MOEA exhibit relatively good performance among the six baseline approaches, whereas they underperform the EMOFM-based approaches that combine both fuzzy clustering and MOEA. Moreover, the Friedman test with the Nemenyi procedure [64] is conducted on the average $Q_{ov}$ of MR-MOEA and the three EMOFM-based approaches, which is a nonparametric statistical procedure for checking whether a set

TABLE V
STATISTICAL RESULTS OF $Q_{ov}$ OBTAINED BY SIX BASELINE APPROACHES AND THREE EMOFM-BASED APPROACHES ON 16 REAL-WORLD NETWORKS

| Network | Metric | LMD | Zhang | NMF | IMOQPSO | MCMOEA | MR-MOEA | EMOFM-SC | EMOFM-RW | EMOFM-DK |
|---|---|---|---|---|---|---|---|---|---|---|
| ENZYMES_g163 | $Q_{ov}\_Max$ | 0.2397 | 0.2438 | 0.1963 | 0.2438 | 0.2438 | **0.2567** | **0.2567** | **0.2567** | **0.2567** |
| | $Q_{ov}\_Avg$ | 0.2397 | 0.2438 | 0.1963 | 0.2431 | 0.2432 | **0.2567** | **0.2567** | **0.2567** | **0.2567** |
| | $Std$ | 0 | 0 | 0 | 0.0023 | 0.0035 | 0 | 0 | 0 | 0 |
| Karate | $Q_{ov}\_Max$ | 0.2162 | 0.1900 | 0.2081 | 0.2130 | 0.2107 | 0.2285 | **0.2348** | **0.2348** | **0.2348** |
| | $Q_{ov}\_Avg$ | 0.2145 | 0.1900 | 0.2081 | 0.2080 | 0.2087 | 0.2261 | **0.2346** | 0.2325 | 0.2341 |
| | $Std$ | 0.0022 | 0 | 0 | 0.0040 | 0.0017 | 0.0039 | 0.0001 | 0.0027 | 0.0008 |
| Dolphin | $Q_{ov}\_Max$ | 0.2614 | 0.2349 | 0.2636 | 0.2088 | 0.2212 | 0.2626 | 0.2718 | **0.2741** | 0.2730 |
| | $Q_{ov}\_Avg$ | 0.2600 | 0.2236 | 0.2636 | 0.2040 | 0.2024 | 0.2597 | 0.2700 | 0.2717 | **0.2723** |
| | $Std$ | 0.0009 | 0.0062 | 0.0000 | 0.0028 | 0.0103 | 0.0022 | 0.0008 | 0.0016 | 0.0008 |
| Polbook | $Q_{ov}\_Max$ | 0.2689 | 0.2544 | 0.2591 | 0.2289 | 0.2379 | 0.2649 | 0.2702 | **0.2704** | **0.2704** |
| | $Q_{ov}\_Avg$ | 0.2554 | 0.2515 | 0.2591 | 0.2269 | 0.2179 | 0.2631 | 0.2701 | 0.2702 | **0.2703** |
| | $Std$ | 0.0054 | 0.002 | 0 | 0.0022 | 0.0136 | 0.0015 | 0.0001 | 0.0002 | 0.0001 |
| Football | $Q_{ov}\_Max$ | 0.2982 | 0.2889 | 0.3049 | 0.2431 | 0.2785 | 0.3049 | **0.3067** | **0.3067** | **0.3067** |
| | $Q_{ov}\_Avg$ | 0.2917 | 0.2874 | 0.3049 | 0.2352 | 0.2726 | 0.3035 | **0.3066** | 0.3063 | **0.3066** |
| | $Std$ | 0.0031 | 0.0019 | 0.0000 | 0.0143 | 0.0064 | 0.0018 | 0.0000 | 0.0003 | 0.0001 |
| Email | $Q_{ov}\_Max$ | 0.2287 | 0.2102 | 0.2368 | 0.1170 | 0.0826 | 0.2463 | 0.2799 | 0.2783 | **0.2823** |
| | $Q_{ov}\_Avg$ | 0.2234 | 0.2033 | 0.2352 | 0.1132 | 0.0774 | 0.2333 | 0.2779 | 0.2723 | **0.2784** |
| | $Std$ | 0.0031 | 0.0025 | 0.0024 | 0.0023 | 0.0029 | 0.0069 | 0.0022 | 0.0031 | 0.0028 |
| Blogs | $Q_{ov}\_Max$ | 0.3591 | 0.3429 | 0.3991 | 0.3561 | 0.3347 | 0.3955 | **0.4173** | 0.4013 | 0.4165 |
| | $Q_{ov}\_Avg$ | 0.3555 | 0.3297 | 0.3967 | 0.3534 | 0.3297 | 0.3875 | **0.4137** | 0.3977 | 0.4131 |
| | $Std$ | 0.0021 | 0.0025 | 0.0023 | 0.0023 | 0.0025 | 0.0027 | 0.0020 | 0.0028 | 0.0018 |
| ego-Facebook | $Q_{ov}\_Max$ | 0.4041 | - | 0.3520 | - | - | 0.4125 | 0.4111 | **0.4166** | 0.4145 |
| | $Q_{ov}\_Avg$ | 0.4029 | - | 0.3492 | - | - | 0.4064 | 0.4090 | **0.4151** | 0.4126 |
| | $Std$ | 0.0013 | - | 0.0028 | - | - | 0.0037 | 0.0021 | 0.0015 | 0.0010 |
| SFI | $Q_{ov}\_Max$ | 0.3496 | 0.3173 | 0.3765 | 0.3283 | 0.2285 | 0.3741 | **0.3851** | 0.3841 | **0.3851** |
| | $Q_{ov}\_Avg$ | 0.3459 | 0.3032 | 0.3765 | 0.3242 | 0.1976 | 0.3672 | 0.3821 | 0.3826 | **0.3843** |
| | $Std$ | 0.0026 | 0.0050 | 0.0000 | 0.0022 | 0.0131 | 0.0040 | 0.0013 | 0.0010 | 0.0010 |
| Jazz | $Q_{ov}\_Max$ | 0.1485 | 0.1482 | 0.1653 | 0.1562 | 0.1458 | 0.2190 | **0.2259** | **0.2259** | **0.2259** |
| | $Q_{ov}\_Avg$ | 0.1465 | 0.1394 | 0.1653 | 0.0873 | 0.1324 | 0.2169 | 0.2258 | **0.2259** | 0.2258 |
| | $Std$ | 0.0008 | 0.0132 | 0.0000 | 0.0514 | 0.0091 | 0.0023 | 0.0002 | 0.0001 | 0.0001 |
| Netscience | $Q_{ov}\_Max$ | 0.4263 | 0.3196 | 0.4655 | 0.3623 | 0.4522 | 0.4612 | 0.4718 | 0.4713 | **0.4733** |
| | $Q_{ov}\_Avg$ | 0.4212 | 0.3195 | 0.4496 | 0.3592 | 0.4482 | 0.4563 | 0.4698 | 0.4696 | **0.4709** |
| | $Std$ | 0.0022 | 0.0001 | 0.0101 | 0.0034 | 0.0022 | 0.0109 | 0.0009 | 0.0012 | 0.0013 |
| Erdös | $Q_{ov}\_Max$ | 0.2824 | - | 0.3126 | - | - | 0.3140 | **0.3349** | 0.3166 | 0.3315 |
| | $Q_{ov}\_Avg$ | 0.2822 | - | 0.3093 | - | - | 0.2970 | **0.3329** | 0.3086 | 0.3283 |
| | $Std$ | 0.0021 | - | 0.0035 | - | - | 0.0108 | 0.0019 | 0.0055 | 0.0019 |
| Celegansmetabolic | $Q_{ov}\_Max$ | 0.1401 | 0.0608 | 0.0968 | 0.0490 | 0.0336 | 0.0778 | **0.2211** | 0.2187 | 0.2187 |
| | $Q_{ov}\_Avg$ | 0.1352 | 0.0594 | 0.0911 | 0.0456 | 0.0284 | 0.0726 | **0.2178** | 0.2154 | 0.2171 |
| | $Std$ | 0.0028 | 0.0015 | 0.0042 | 0.0018 | 0.0039 | 0.0026 | 0.0020 | 0.0023 | 0.0013 |
| Yeast-D2 | $Q_{ov}\_Max$ | 0.3972 | 0.3912 | 0.2221 | 0.3020 | 0.3502 | 0.4132 | 0.4147 | 0.4144 | **0.4193** |
| | $Q_{ov}\_Avg$ | 0.3944 | 0.3893 | 0.2100 | 0.2541 | 0.3351 | 0.4081 | 0.4139 | 0.4115 | **0.4179** |
| | $Std$ | 0.0015 | 0.0071 | 0.0067 | 0.0392 | 0.0100 | 0.0037 | 0.0013 | 0.0019 | 0.0010 |
| Y2H | $Q_{ov}\_Max$ | 0.2970 | 0.2519 | 0.3098 | 0.2810 | 0.2362 | 0.3223 | 0.3523 | 0.3455 | **0.3661** |
| | $Q_{ov}\_Avg$ | 0.2943 | 0.2508 | 0.3085 | 0.2810 | 0.2274 | 0.3152 | 0.3483 | 0.3423 | **0.3628** |
| | $Std$ | 0.0008 | 0.0004 | 0.008 | 0 | 0.005 | 0.0055 | 0.0020 | 0.0021 | 0.0028 |
| PPI | $Q_{ov}\_Max$ | 0.2911 | 0.2569 | 0.3303 | 0.2562 | 0.2072 | 0.3235 | **0.3500** | 0.3377 | 0.3482 |
| | $Q_{ov}\_Avg$ | 0.2875 | 0.2561 | 0.3278 | 0.2510 | 0.1990 | 0.3193 | **0.3468** | 0.3304 | 0.3460 |
| | $Std$ | 0.0017 | 0.0006 | 0.0014 | 0.0052 | 0.0053 | 0.0036 | 0.0023 | 0.0037 | 0.0017 |

The best result in each row is shown in bold font.

TABLE VI
FRIEDMAN TEST WITH NEMENYI PROCEDURE ON AVERAGE $Q_{ov}$ OBTAINED BY MR-MOEA AND THREE EMOFM-BASED APPROACHES, WHERE "1" INDICATES SIGNIFICANT DIFFERENCE AND "0" OTHERWISE

| Algorithm | MR-MOEA | EMOFM-RW | EMOFM-SC | EMOFM-DK |
|---|---|---|---|---|
| MR-MOEA | NaN | 1 | 1 | 1 |
| EMOFM-RW | 1 | NaN | 0 | 0 |
| EMOFM-SC | 1 | 0 | NaN | 0 |
| EMOFM-DK | 1 | 0 | 0 | NaN |
| Friedman rank | 1.0938 | 2.4688 | 3.0313 | 3.4063 |
| $p$-value | 5.2679e-07 | significance level $\alpha$ | | 0.05 |

of samples are statistically different. According to Table VI, the $p$-value smaller than the significance level $\alpha$ indicates that the performance of the four approaches is significantly different, where MR-MOEA is statistically different from the EMOFM-based approaches, and the EMOFM-based approaches are statistically similar to each other. Besides, the Friedman ranks indicate that MR-MOEA has worse performance than the EMOFM-based approaches, since a smaller Friedman rank indicates a smaller $Q_{ov}$. As a result, the superiority of the proposed EMOFM over existing approaches for overlapping community detection is confirmed.

Table VII lists the gNMI values obtained by EMOFM and the compared approaches on the real-world networks whose ground truth is known.[1] It can be observed from the table that the EMOFM-based approaches also perform the best on five out of six networks, while their gNMI values are slightly smaller than those of NMF, IMOQPSO, and MR-MOEA on the Polbook network. This is due to the inconsistency between the metrics $Q_{ov}$ and gNMI [65], and the EMOFM-based approaches can obtain better results in terms of $Q_{ov}$ since they directly adopt $Q_{ov}$ as an objective to be optimized.

To take a close look at the results obtained by different approaches, Table VIII presents the numbers of communities and overlapping nodes found by EMOFM and the compared approaches, where each cell shows the minimum and maximum numbers of communities or overlapping nodes found by an approach in 30 runs. Owing to the population-based

[1]The ground truths of Karate, Dolphin, Polbook, and Football networks can be obtained from http://web.xidian.edu.cn/rhshang/paper.html, and the ground truths of Yeast-D2 and Y2H networks can be obtained from [Online]. Available: http://faculty.uaeu.ac.ae/nzaki/ProRank.htm.

TABLE VII
STATISTICAL RESULTS OF gNMI OBTAINED BY SIX BASELINE APPROACHES AND THREE EMOFM-BASED APPROACHES ON 6 REAL-WORLD NETWORKS

| Network | Metric | LMD | Zhang | NMF | IMOQPSO | MCMOEA | MR-MOEA | EMOFM-SC | EMOFM-RW | EMOFM-DK |
|---|---|---|---|---|---|---|---|---|---|---|
| Karate | $gNMI\_Max$ | 0.5135 | 0.3477 | 0.4404 | 0.7081 | 0.9186 | **1** | **1** | **1** | **1** |
| | $gNMI\_Avg$ | 0.398 | 0.3477 | 0.4404 | 0.6982 | 0.8044 | 0.9837 | **1** | **1** | **1** |
| | $Std$ | 0.0741 | 0 | 0 | 0.0241 | 0.1194 | 0.0517 | 0.0000 | 0.0000 | 0.0000 |
| Dolphin | $gNMI\_Max$ | 0.6110 | 0.3546 | 0.4664 | 0.7944 | 0.3134 | **1** | 0.8889 | **1** | **1** |
| | $gNMI\_Avg$ | 0.4562 | 0.3456 | 0.4664 | 0.7931 | 0.2734 | 0.9667 | 0.8889 | 0.9667 | **1** |
| | $Std$ | 0.1321 | 0.0138 | 0.0000 | 0.0042 | 0.0277 | 0.0537 | 0.0000 | 0.0537 | 0.0000 |
| Polbook | $gNMI\_Max$ | 0.2295 | 0.1790 | 0.3880 | 0.4331 | 0.3079 | **0.4587** | 0.3496 | 0.3514 | 0.3371 |
| | $gNMI\_Avg$ | 0.2042 | 0.1783 | 0.3880 | 0.4065 | 0.2742 | **0.4228** | 0.3495 | 0.3395 | 0.3358 |
| | $Std$ | 0.0311 | 0.0005 | 0 | 0.0193 | 0.0233 | 0.0407 | 0.0001 | 0.0063 | 0.0009 |
| Football | $gNMI\_Max$ | 0.7702 | 0.7119 | 0.8032 | 0.8092 | 0.7282 | 0.8030 | 0.8524 | **0.8565** | 0.8376 |
| | $gNMI\_Avg$ | 0.7230 | 0.7082 | 0.8032 | 0.7984 | 0.7097 | 0.8030 | 0.8129 | **0.8285** | 0.8245 |
| | $Std$ | 0.0306 | 0.0048 | 0.0000 | 0.0153 | 0.0157 | 0 | 0.0179 | 0.0164 | 0.0130 |
| Yeast-D2 | $gNMI\_Max$ | 0.2117 | 0.1844 | 0.2078 | 0.2052 | 0.2658 | 0.2641 | 0.2669 | 0.2625 | **0.5401** |
| | $gNMI\_Avg$ | 0.2050 | 0.1824 | 0.1968 | 0.2013 | 0.2548 | 0.2542 | 0.2667 | 0.2587 | **0.3162** |
| | $Std$ | 0.0034 | 0.0013 | 0.0066 | 0.0041 | 0.0065 | 0.0047 | 0.0004 | 0.0037 | 0.1180 |
| Y2H | $gNMI\_Max$ | 0.0183 | 0.0313 | 0.0261 | 0.0184 | 0.0273 | 0.1182 | 0.1061 | **0.1383** | 0.0885 |
| | $gNMI\_Avg$ | 0.0183 | 0.0313 | 0.0261 | 0.0183 | 0.0227 | 0.1154 | 0.1006 | **0.1178** | 0.0856 |
| | $Std$ | 0 | 0 | 0 | 0.001 | 0.0033 | 0.0019 | 0.0032 | 0.0102 | 0.0014 |

The best result in each row is shown in bold font.

TABLE VIII
MINIMUM NUMBER OF COMMUNITIES $C\_\mathrm{Min}$, MAXIMUM NUMBER OF COMMUNITIES $C\_\mathrm{Max}$, MINIMUM NUMBER OF OVERLAPPING NODES $O\_\mathrm{Min}$, AND MAXIMUM NUMBER OF OVERLAPPING NODES $O\_\mathrm{Max}$ FOUND BY SIX BASELINE APPROACHES AND THREE EMOFM-BASED APPROACHES ON 16 REAL-WORLD NETWORKS

| Network | Metric | LMD | Zhang | NMF | IMOQPSO | MCMOEA | MR-MOEA | EMOFM-SC | EMOFM-RW | EMOFM-DK |
|---|---|---|---|---|---|---|---|---|---|---|
| ENZYMES_g163 | $(C\_Min, C\_Max)$ | (2,2) | (2,2) | (3,3) | (1,6) | (2,4) | (1,12) | (1,12) | (1,12) | (1,12) |
| | $(O\_Min, O\_Max)$ | (0,0) | (1,1) | (4,4) | (0,10) | (1,5) | (0,1) | (0,6) | (0,6) | (0,6) |
| Karate | $(C\_Min, C\_Max)$ | (3,4) | (4,4) | (4,4) | (3,7) | (2,6) | (1,32) | (1,30) | (1,29) | (1,28) |
| | $(O\_Min, O\_Max)$ | (3,4) | (8,8) | (8,8) | (2,31) | (1,10) | (0,3) | (0,26) | (0,24) | (0,25) |
| Dolphin | $(C\_Min, C\_Max)$ | (3,6) | (6,8) | (2,2) | (2,7) | (6,17) | (2,49) | (1,35) | (1,36) | (1,38) |
| | $(O\_Min, O\_Max)$ | (5,12) | (13,20) | (2,2) | (7,34) | (15,34) | (0,13) | (0,42) | (0,47) | (0,45) |
| Polbook | $(C\_Min, C\_Max)$ | (2,6) | (6,6) | (12,12) | (1,7) | (4,14) | (3,95) | (1,51) | (1,48) | (1,55) |
| | $(O\_Min, O\_Max)$ | (1,16) | (17,19) | (61,61) | (0,92) | (14,58) | (0,11) | (0,86) | (0,81) | (0,83) |
| Football | $(C\_Min, C\_Max)$ | (8,12) | (11,11) | (12,12) | (3,9) | (12,18) | (2,97) | (1,67) | (1,63) | (1,64) |
| | $(O\_Min, O\_Max)$ | (4,17) | (12,14) | (1,1) | (46,85) | (1,21) | (4,16) | (0,95) | (0,100) | (0,97) |
| Email | $(C\_Min, C\_Max)$ | (40,52) | (60,66) | (12,12) | (2,7) | (228,282) | (8,895) | (1,348) | (2,324) | (2,377) |
| | $(O\_Min, O\_Max)$ | (321,405) | (518,566) | (343,385) | (4,700) | (608,661) | (95,145) | (0,906) | (0,1109) | (0,1046) |
| Blogs | $(C\_Min, C\_Max)$ | (346,357) | (361,366) | (10,10) | (3,7) | (366,439) | (213,3480) | (2,968) | (1,961) | (2,938) |
| | $(O\_Min, O\_Max)$ | (461,531) | (761,804) | (252,296) | (200,384) | (690,746) | (217,298) | (0,3152) | (0,3455) | (0,1839) |
| ego-Facebook | $(C\_Min, C\_Max)$ | (59,65) | - | (117,156) | - | - | (8,4001) | (2,860) | (2,651) | (2,961) |
| | $(O\_Min, O\_Max)$ | (189,307) | - | (249,351) | - | - | (13,31) | (0,3100) | (0,3986) | (0,3964) |
| SFI | $(C\_Min, C\_Max)$ | (12,14) | (12,12) | (16,16) | (1,6) | (25,43) | (9,109) | (1,59) | (1,57) | (1,61) |
| | $(O\_Min, O\_Max)$ | (2,8) | (10,19) | (14,14) | (0,14) | (22,37) | (1,9) | (0,56) | (0,71) | (0,50) |
| Jazz | $(C\_Min, C\_Max)$ | (2,3) | (5,5) | (4,4) | (2,7) | (5,17) | (2,158) | (1,75) | (1,67) | (1,78) |
| | $(O\_Min, O\_Max)$ | (0,25) | (4,7) | (5,5) | (25,179) | (20,103) | (10,29) | (0,189) | (0,194) | (0,190) |
| Netscience | $(C\_Min, C\_Max)$ | (437,446) | (572,572) | (12,12) | (2,4) | (303,335) | (305,1376) | (2,413) | (2,414) | (2,396) |
| | $(O\_Min, O\_Max)$ | (20,36) | (388,389) | (3,12) | (2,1461) | (45,79) | (27,56) | (0,1336) | (0,944) | (0,646) |
| Erdös | $(C\_Min, C\_Max)$ | (380,386) | - | (22,39) | - | - | (52,6324) | (2,1805) | (2,1705) | (2,1732) |
| | $(O\_Min, O\_Max)$ | (1233,1291) | - | (774,909) | - | - | (264,361) | (0,5740) | (0,6610) | (0,4661) |
| Celegans metabolic | $(C\_Min, C\_Max)$ | (9,16) | (7,9) | (11,11) | (2,9) | (246,274) | (196,359) | (1,140) | (1,102) | (1,153) |
| | $(O\_Min, O\_Max)$ | (0,91) | (197,210) | (40,41) | (215,382) | (116,137) | (31,62) | (0,319) | (0,439) | (0,430) |
| Yeast-D2 | $(C\_Min, C\_Max)$ | (143,152) | (213,218) | (172,186) | (3,5) | (155,196) | (106,1313) | (2,400) | (2,357) | (1,410) |
| | $(O\_Min, O\_Max)$ | (48,83) | (693,737) | (462,509) | (7,145) | (207,317) | (48,87) | (0,1229) | (0,1208) | (0,1018) |
| Y2H | $(C\_Min, C\_Max)$ | (447,458) | (502,507) | (203,203) | (4,4) | (418,482) | (244,1466) | (2,612) | (2,520) | (2,529) |
| | $(O\_Min, O\_Max)$ | (223,246) | (516,533) | (261,284) | (78,1966) | (554,611) | (216,298) | (0,1325) | (0,1712) | (0,1132) |
| PPI | $(C\_Min, C\_Max)$ | (202,217) | (251,257) | (237,245) | (3,6) | (355,422) | (79,1923) | (2,659) | (2,630) | (2,681) |
| | $(O\_Min, O\_Max)$ | (475,551) | (884,917) | (310,324) | (30,563) | (909,992) | (222,310) | (0.1993) | (0.2375) | (0,1520) |

search mechanism of MOEA-based approaches, the numbers of communities and overlapping nodes found by them have much larger ranges than those found by LMD, Zhang's approach, and NMF. As for the EMOFM-based approaches, the numbers of communities found by them have larger ranges than those found by the other approaches besides MR-MOEA and the numbers of overlapping nodes found by them have larger ranges than those found by all the other approaches. As a result, the proposed EMOFM can find a number of diverse community structures with different numbers of communities and overlapping nodes.

According to Tables V and VII, although none of the three EMOFM-based approaches can obtain the best results on all the real-world networks, they outperform the six baseline approaches in most cases. Moreover, their metric values are very similar to each other, which indicates that the performance of the proposed EMOFM is not sensitive to the distance measure used. For simplicity, we use EMOFM-DK as a representative of the proposed EMOFM in the subsequent experiments.

### C. Results on LFR Benchmark Networks

The average gNMI values obtained by six baseline approaches and EMOFM-DK on 72 LFR networks are shown in Fig. 5, where two observations can be made. First, EMOFM-DK outperforms the other approaches when the mixing parameter $\mu$ is small. This is because a small $\mu$ indicates an unambiguous community structure, which enables EMOFM-DK to find the central nodes easily. Second, EMOFM-DK is superior over the other approaches, especially when the overlapping membership $O_m$ is large, since EMOFM-DK assigns a fuzzy threshold to each node, which is helpful to the detection of nodes belonging to multiple communities. As a consequence, the proposed EMOFM also exhibits the best overall performance on the LFR networks considered here.

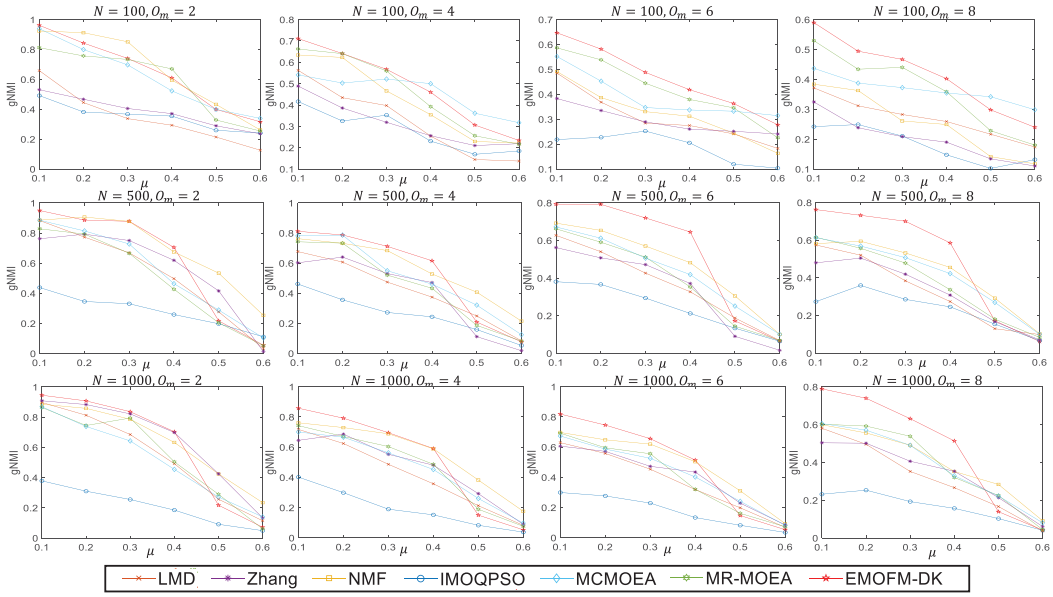


Fig. 5. Average gNMI obtained by six baseline approaches and EMOFM-DK on 72 LFR networks.

TABLE IX
AVERAGE $Q_{ov}$ OBTAINED BY THE FIRST STAGES OF EMOFM-DK AND EMOFM-DK WITH RANDOM INITIALIZATION AND THREE MOEA-BASED APPROACHES ON 14 REAL-WORLD NETWORKS

| Network | First stage of EMOFM-DK | First stage of EMOFM-DK with random initialization | IMOQPSO | MCMOEA | MR-MOEA |
|---|---|---|---|---|---|
| ENZYMES_g163 | **0.2567** | **0.2567** | 0.2431 | 0.2432 | **0.2567** |
| Karate | **0.2315** | 0.2258 | 0.2080 | 0.2087 | 0.2261 |
| Dolphin | **0.2634** | 0.2359 | 0.2040 | 0.2024 | 0.2597 |
| Polbook | **0.2667** | 0.1836 | 0.2269 | 0.2179 | 0.2631 |
| Football | 0.2948 | 0.1861 | 0.2352 | 0.2726 | **0.3035** |
| Email | **0.2539** | 0.0856 | 0.1132 | 0.0774 | 0.2333 |
| Blogs | 0.3872 | 0.1823 | 0.3534 | 0.3297 | **0.3875** |
| SFI | **0.3741** | 0.3054 | 0.3242 | 0.1976 | 0.3672 |
| Jazz | 0.2148 | 0.0479 | 0.0873 | 0.1324 | **0.2169** |
| Netscience | 0.4487 | 0.2437 | 0.3592 | 0.4482 | **0.4563** |
| Celegans metabolic | **0.2005** | 0.1205 | 0.0456 | 0.0284 | 0.0726 |
| Yeast-D2 | 0.4008 | 0.1096 | 0.2541 | 0.3351 | **0.4081** |
| Y2H | **0.3380** | 0.1906 | 0.2810 | 0.2274 | 0.3152 |
| PPI | **0.3247** | 0.1409 | 0.2510 | 0.1990 | 0.3193 |

The best result in each row is shown in bold font.

### D. Effectiveness of the Population Initialization Strategy in EMOFM

We design two population initialization strategies in the proposed EMOFM, where the initialization strategy in the first stage can improve the quality of initial individuals, and the initialization strategy in the second strategy is necessary for generating a subpopulation for each individual. Since the initialization strategy in the first stage can be removed from EMOFM, we verify the effectiveness of the initialization strategy in the first stage by comparing it with the general initialization strategy, i.e., random initialization.

The average $Q_{ov}$ values obtained by the first stages (i.e., finding only nonoverlapping communities) of EMOFM-DK and EMOFM-DK with random initialization on 14 real-world networks can be found in Table IX. As can be observed, EMOFM-DK outperforms EMOFM-DK with random initialization on all the networks. For further observations, Fig. 6

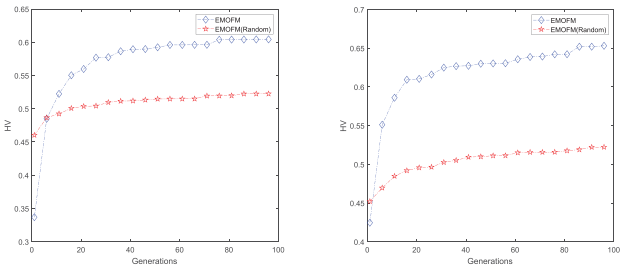


Fig. 6. Convergence profiles of HV obtained by EMOFM-DK and EMOFM-DK with random initialization strategy on two benchmark networks. (a) Email network. (b) Yeast-D2 network.

depicts the convergence profiles of hypervolume (HV) [66] obtained by the two approaches on the Email network and the Yeast-D2 network, averaged over 30 runs. Given a population, the HV value is defined as the area covered by all the individuals in the population with respect to a reference point in objective space, where a larger HV value indicates a better convergence and diversity of the population. As shown in Fig. 6, the population generated by EMOFM-DK is worse than that generated by EMOFM-DK with random initialization in terms of HV; however, the former leads to a much faster convergence speed than the latter. Therefore, the proposed initialization strategy can significantly improve the performance of EMOFM.

### E. Effectiveness of the Two-Stage Optimization Process in EMOFM

To verify the effectiveness of the nonoverlapping community detection in the first stage of EMOFM, we compare the average $Q_{ov}$ values obtained by the first stage of EMOFM-DK with those obtained by IMOQPSO, MCMOEA, and MR-MOEA. According to the results presented in Table IX, it can be found that the first stage of EMOFM-DK outperforms the other

TABLE X
FRIEDMAN TEST WITH NEMENYI PROCEDURE ON AVERAGE $Q_{ov}$ OBTAINED BY THE FIRST STAGES OF EMOFM-DK, EMOFM-DK WITH RANDOM INITIALIZATION, AND THREE MOEA-BASED APPROACHES, WHERE "1" INDICATES SIGNIFICANT DIFFERENCE AND "0" OTHERWISE

| Algorithm | First stage of EMOFM-DK | First stage of EMOFM-DK with random initialization | IMOQPSO | MCMOEA | MR-MOEA |
|---|---|---|---|---|---|
| First stage of EMOFM-DK | NaN | 1 | 1 | 1 | 0 |
| First stage of EMOFM-DK with random initialization | 1 | NaN | 0 | 0 | 1 |
| IMOQPSO | 1 | 0 | NaN | 0 | 1 |
| MCMOEA | 1 | 0 | 0 | NaN | 1 |
| MR-MOEA | 0 | 1 | 1 | 1 | NaN |
| Friedman rank | 4.6071 | 1.8929 | 2.2857 | 2.0000 | 4.2143 |
| $p$-value | 1.0291e-07 | | significance level $\alpha$ | | 0.05 |

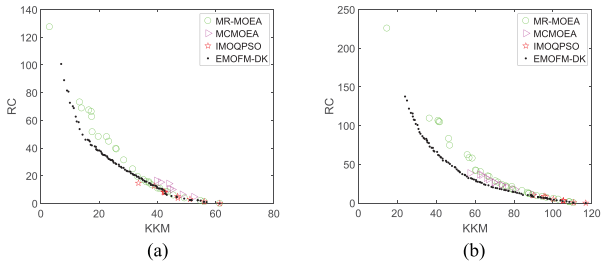


Fig. 7. Populations obtained by IMOQPSO, MCMOEA, MR-MOEA, and the first stage of EMOFM-DK on two benchmark networks. (a) Karate network. (b) Dolphin network.

MOEA-based approaches on 8 out of 14 real-world networks. Besides, Table X presents the statistical results obtained by Friedman test with the Nemenyi procedure, it can be seen that the performance of the first stage of EMOFM-DK is statistically different from the performance of IMOQPSO and MCMOEA, and statistically similar to the performance of MR-MOEA. Furthermore, Fig. 7 plots the populations with the median HV value among 30 runs obtained by the four MOEA-based approaches on the Karate network and the Dolphin network. As can be seen from Fig. 7, the community structures obtained in the first stage of EMOFM-DK have significantly better KKM and RC values than those obtained by the other MOEA-based approaches. In fact, the inferiority of the other approaches is mainly attributed to the prior knowledge they rely on, which may lead to unexpected results if the prior knowledge is not ideal.

Last, we verify the effectiveness of the overlapping community detection in the second stage of EMOFM. Figs. 8 and 9 depict the overlapping communities obtained by EMOFM-DK, LMD, Zhang's approach, and NMF on the ENZYMES_g163 network and the Karate network, respectively. As can be observed, the overlapping communities found by EMOFM-DK shown in Fig. 8(a) is the same to that found by Zhang's approach, and the overlapping communities found by EMOFM-DK shown in Fig. 8(b) is with better $Q_{ov}$; besides, EMOFM-DK can find more overlapping nodes as shown in Fig. 8(c)–(d). Similarly, the overlapping communities found by EMOFM-DK shown in Fig. 9(a)–(c) are with better $Q_{ov}$ than those found by LMD, Zhang's approach, and NMF. To summarize, the proposed EMOFM can not only find the overlapping communities similar to those detected by the other approaches, but also detect more

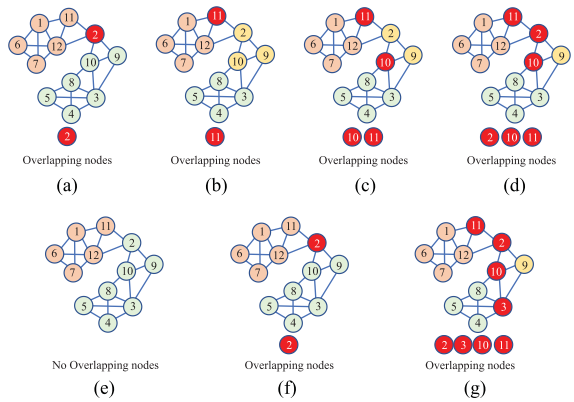


Fig. 8. Four overlapping community structures obtained by EMOFM-DK and the overlapping communities obtained by LMD, Zhang's approach, and NMF on the ENZYMES_g163 network. (a) EMOFM-DK ($Q_{ov} = 0.2438$). (b) EMOFM-DK ($Q_{ov} = 0.2485$). (c) EMOFM-DK ($Q_{ov} = 0.2417$). (d) EMOFM-DK ($Q_{ov} = 0.2257$). (e) LMD ($Q_{ov} = 0.2397$). (f) Zhang ($Q_{ov} = 0.2438$). (g) NMF ($Q_{ov} = 0.1963$).

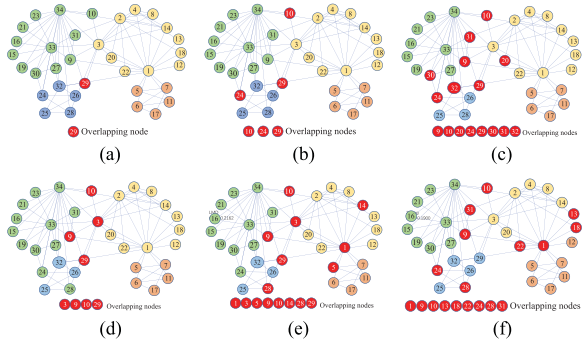


Fig. 9. Three overlapping community structures obtained by EMOFM-DK and the overlapping communities obtained by LMD, Zhang's approach, and NMF on the Karate network. (a) EMOFM-DK ($Q_{ov} = 0.2326$). (b) EMOFM-DK ($Q_{ov} = 0.2317$). (c) EMOFM-DK ($Q_{ov} = 0.2092$). (d) LMD ($Q_{ov} = 0.2162$). (e) Zhang ($Q_{ov} = 0.1900$). (f) NMF ($Q_{ov} = 0.2081$).

promising overlapping communities with different overlapping degrees.

### F. Discussion

The abovementioned experimental results have demonstrated the effectiveness of the proposed EMOFM on overlapping community detection, while they also reveal the limitations of EMOFM. First, according to the results on LFR benchmark networks shown in Fig. 5, the performance of EMOFM deteriorates considerably when the community structure is not distinct, since the central nodes of ambiguous communities are hard to be detected, and thus, the membership among nodes is inaccurate. Second, according to Fig. 8(g), node 3 is likely to be an overlapping node, but it is missed by EMOFM since it is detected as a central node and can never be an overlapping node due to the representation of EMOFM. Besides, according to the analysis in Section III-E, the time complexity of EMOFM increases considerably with the increase of number of nodes $n$, which is a common limitation of MOEA-based approaches. Table XI lists the average runtime of EMOFM-DK and the other MOEA-based approaches on 16 real-world networks, it can be found that the

TABLE XI
AVERAGE RUNTIME OF EMOFM-DK AND THREE MOEA-BASED APPROACHES ON 16 REAL-WORLD NETWORKS

| Network | Nodes | Runtime | | | |
|---|---|---|---|---|---|
| | | IMOQPSO | MCMOEA | MR-MOEA | EMOFM-DK |
| ENZYMES_g163 | 12 | 4.42s | 0.72s | 8.63s | 2.55s |
| Karate | 34 | 5.32s | 1.21s | 19.34s | 3.62s |
| Dolphin | 62 | 6.96s | 2.6s | 56.70s | 5.32s |
| Polbook | 105 | 7.86s | 4.07s | 77.12s | 8.13s |
| Football | 115 | 8.71s | 4.52s | 93.72s | 9.26s |
| SFI | 118 | 8.46s | 4.31s | 66.48s | 9.19s |
| Jazz | 198 | 366.61s | 13s | 173.72s | 16.69s |
| Celegansmetabolic | 453 | 186.67s | 39.90s | 637.92s | 48.16s |
| Email | 1133 | 773.02s | 177.62s | 1529.96s | 299.05s |
| Yeast-D2 | 1443 | 1895.67 | 282.61s | 1805.83s | 605.15s |
| Netscience | 1589 | 538.63s | 108.37s | 1779.51s | 488.78s |
| Y2H | 1966 | 638.56s | 506.94s | 6647.14s | 1238.60s |
| PPI | 2445 | 1505.99s | 944.05s | 8069.49s | 1862.29s |
| Blogs | 3984 | 2758.92s | 979.38s | 15624.02s | 3230.36s |
| ego-Facebook | 4039 | - | - | 5953.07s | 4377.45s |
| Erdös | 6927 | - | - | 26287.01s | 10937.85s |

runtimes on the Erdös network with 6927 nodes is much more than that on the ENZYMES_g163 network with 12 nodes.

## V. CONCLUSION

In this article, we have proposed an EMOFM for overlapping community detection. The proposed EMOFM enhanced the performance of fuzzy clustering by optimizing the community centers via a well-tailored MOEA, and it also optimized the fuzzy thresholds for obtaining diverse overlapping community structures. Furthermore, to improve the performance of the MOEA in optimizing community centers and fuzzy thresholds, a two-stage optimization process and two population initialization strategies were proposed for EMOFM.

In the experiments, we have implemented three approaches by embedding three distance measures in EMOFM. The statistical results have indicated that EMOFM outperforms six state-of-the-art approaches for overlapping community detection, including fuzzy-clustering-based approaches and MOEA-based approaches. In addition, the effectiveness of the population initialization strategy and the two-stage optimization process in EMOFM have also been verified.

This article has demonstrated the potential of combing fuzzy clustering and MOEAs on overlapping community detection, and we would like to further explore the proposed EMOFM from the following aspects: First, it is desirable to investigate other distance measures or an ensemble of multiple distance measures to further improve the performance of EMOFM. Second, due to the resolution limit in community detection [67], some objectives for measuring the quality of overlapping communities other than $Q_{ov}$ can be adopted. Third, some more efficient search strategies can be developed to reduce the time complexity of EMOFM.

## REFERENCES


[1] C. Pizzuti and S. E. Rombo, "Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods," *Bioinformatics*, vol. 30, no. 10, pp. 1343–52, 2014.

[2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[4] Q. Cai, L. Ma, and M. Gong, "A survey on network community detection based on evolutionary computation," *Int. J. Bio-Inspired Comput.*, vol. 8, no. 2, pp. 84–98, 2014.

[5] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Rev. E*, vol. 74, no. 1, pp. 016–110, 2006.

[6] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Rev. E Statistical Nonlinear Soft Matter Phys.*, vol. 83, no. 1, 2011, Art. no. 016107.

[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[8] Q. Chen, T.-T. Wu, and M. Fang, "Detecting local community structures in complex networks based on local degree central nodes," *Physica A Statistical Mech. Appl.*, vol. 392, no. 3, pp. 529–537, 2013.

[9] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, pp. 2011–2024, 2010.

[10] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[11] X. Wang, G. Liu, L. Pan, and J. Li, "Uncovering fuzzy communities in networks with structural similarity," *Neurocomputing*, vol. 210, pp. 26–33, 2016.

[12] L. Zhang, H. Pan, Y. Su, X. Zhang, and Y. Niu, "A mixed representation-based multiobjective evolutionary algorithm for overlapping community detection," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2703–2716, Sep. 2017.

[13] L. Hu and K. C. C. Chan, "Fuzzy clustering in a complex network based on content relevance and link structures," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 2, pp. 456–470, Apr. 2016.

[14] S. Zhang, R. S. Wang, and X. S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering," *Physica A Statistical Mech. Appl.*, vol. 374, no. 1, pp. 483–490, 2007.

[15] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy *c*-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2, pp. 191–203, 1984.

[16] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó, "Fuzzy communities and the concept of bridgeness in complex networks," *Physical Rev. E*, vol. 77, no. 1, 2008, Art. no. 016107.

[17] W. Wang, D. Liu, X. Liu, and L. Pan, "Fuzzy overlapping community detection based on local random walk and multidimensional scaling," *Physica A Statistical Mech. Appl.*, vol. 392, no. 24, pp. 6578–6586, 2013.

[18] S. Wikaisuksakul, "A multi-objective genetic algorithm with fuzzy *c*-means for automatic data clustering," *Appl. Soft Comput.*, vol. 24, pp. 679–691, 2014.

[19] L. Wang, Y. Liu, X. Zhao, and Y. Xu, "Particle swarm optimization for fuzzy c-means clustering," in *Proc. IEEE World Congr. Intell. Control Autom.*, 2006, vol. 2, pp. 6055–6058.

[20] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1835–1838, 2011.

[21] Z. Ding, X. Zhang, D. Sun, and B. Luo, "Overlapping community detection based on network decomposition," *Scientific Rep.*, vol. 6, 2016, Art. no. 24115.

[22] A. Biswas and B. Biswas, "FuzAg: Fuzzy agglomerative community detection," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2568–2577, Oct. 2018.

[23] T. C. Havens, J. C. Bezdek, C. Leckie, K. Ramamohanarao, and M. Palaniswami, "A soft modularity function for detecting fuzzy communities in social networks," *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 6, pp. 1170–1175, Dec. 2013.

[24] Y. Tian, R. Cheng, X. Zhang, F. Cheng, and Y. Jin, "An indicator based multi-objective evolutionary algorithm with reference point adaptation for better versatility," *IEEE Trans. Evol. Comput.*, vol. 22, no. 4, pp. 609–622, Aug. 2018.

[25] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, "Low-complexity fuzzy relational clustering algorithms for web mining," *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 4, pp. 595–607, Aug. 2001.

[26] L. Angelini, S. Boccaletti, D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Identification of network modules by optimization of ratio association," *Chaos: An Interdisciplinary J. Nonlinear Sci.*, vol. 17, no. 2, 2007, Art. no. 023114.

[27] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 18, no. 1, pp. 82–97, Feb. 2014.

[28] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *J. Statistical Mech. Theory Exp.*, vol. 2009, no. 3, pp. 3166–3168, 2008.

[29] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Uncovering fuzzy community structure in complex networks," *Physical Rev. E*, vol. 76, no. 4, 2007, Art. no. 046103.
[30] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, no. 1, pp. 1457–1469, 2004.
[31] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, "Overlapping community detection using Bayesian non-negative matrix factorization," *Phys. Rev. E Statistical Nonlinear Soft Matter Phys.*, vol. 83, no. 2, 2011, Art. no. 066114.
[32] N. Binesh and M. Rezghi, "Fuzzy clustering in community detection based on nonnegative matrix factorization with two novel evaluation criteria," *Appl. Soft Comput.*, vol. 69, pp. 689–703, 2018.
[33] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.
[34] H. L. Capitaine and C. Frélicot, "A cluster-validity index combining an overlap measure and a separation measure based on fuzzy-aggregation operators," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 3, pp. 580–588, Jun. 2011.
[35] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, "Multiobjective evolutionary algorithms: A survey of the state of the art," *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 32–49, Jun. 2011.
[36] I. Saha, U. Maulik, and D. Plewczynski, "A new multi-objective technique for differential fuzzy clustering," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2765–2776, 2011.
[37] Y. Ding and X. Fu, "Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm," *Neurocomputing*, vol. 188, pp. 233–238, 2016.
[38] P. S. Esfahani and A. Saghaei, "A multi-objective approach to fuzzy clustering using ITLBO algorithm," *J. AI Data Mining*, vol. 5, no. 2, pp. 307–317, 2017.
[39] F. Zhao, J. Fan, H. Liu, R. Lan, and C. Chen, "Noise robust multi-objective evolutionary clustering image segmentation motivated by intuitionistic fuzzy information," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 2, pp. 387–401, Feb. 2019.
[40] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015. [Online]. Available: http://networkrepository.com
[41] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[42] S. White and P. Smyth, "A spectral clustering approach to finding communities in graph," in *Proc. SIAM Int. Conf. Data Mining*, 2005, pp. 274–285.
[43] L. Hagen and A. B. Kahng, "A new approach to effective circuit clustering," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 1992, pp. 422–427.
[44] L. Yen, D. Vanvyve, F. Wouters, F. Fouss, M. Verleysen, and M. Saerens, "Clustering using a random walk based distance measure," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2005, pp. 317–324.
[45] D. J. Klein and M. Randić, "Resistance distance," *J. Math. Chemistry*, vol. 12, no. 1, pp. 81–95, 1993.
[46] A. Firat, S. Chatterjee, and M. Yilmaz, "Genetic clustering of social networks using random walks," *Comput. Statist. Data Anal.*, vol. 51, no. 12, pp. 6285–6294, 2007.
[47] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "A novel way of computing dissimilarities between nodes of a graph, with application to collaborative filtering and subspace projection of the graph nodes," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2005, pp. 550–556.
[48] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proc. Int. Conf. Mach. Learn.*, 2002, vol. 2002, pp. 315–322.
[49] T. Zhang and B. Wu, "A method for local community detection by finding core nodes," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 1171–1176.
[50] J. Yang and J. Leskovec, "Overlapping community detection at scale: A nonnegative matrix factorization approach," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2013, pp. 587–596.
[51] Y. Li, Y. Wang, J. Chen, L. Jiao, and R. Shang, "Overlapping community detection through an improved multi-objective quantum-behaved particle swarm optimization," *J. Heuristics*, vol. 21, no. 4, pp. 549–575, 2015.
[52] X. Wen *et al.*, "A maximal clique based multiobjective evolutionary algorithm for overlapping community detection," *IEEE Trans. Evol. Comput.*, vol. 21, no. 3, pp. 363–377, Jun. 2017.
[53] K. Deb and R. B. Agrawal, "Simulated binary crossover for continuous search space," *Complex Syst.*, vol. 9, no. 4, pp. 115–148, 1995.
[54] K. Deb and M. Goyal, "A combined genetic adaptive search (GeneAS) for engineering design," *Comput. Sci. Informat.*, vol. 26, no. 4, pp. 30–45, 1996.
[55] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropological Res.*, vol. 33, no. 4, pp. 452–473, 1977.
[56] D. Lusseau, "The emergent properties of a dolphin social network," in *Proc. Roy. Soc. London B, Biological Sci.*, 2003, pp. S186–S188.
[57] M. E. Newman, "Modularity and community structure in networks," in *Proc. Nat. Acad. Sci.*, 2006, pp. 8577–8582.
[58] P. M. Gleiser and L. Danon, "Community structure in jazz," *Adv. Complex Syst.*, vol. 6, no. 04, pp. 565–573, 2003.
[59] S. He *et al.*, "Cooperative co-evolutionary module identification with application to cancer disease module discovery," *IEEE Trans. Evol. Comput.*, vol. 20, no. 6, pp. 874–891, Dec. 2016.
[60] N. Zaki, J. Berengueres, and D. Efimov, "ProRank: A method for detecting protein complexes," in *Proc. ACM Int. Conf. Genetic Evol. Comput.*, 2012, pp. 209–216.
[61] H. Yu *et al.*, "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–110, 2008.
[62] L. Andrea and F. Santo, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Rev. E Statistical Nonlinear Soft Matter Phys.*, vol. 80, no. 1, pp. 016–118, 2009.
[63] A. Lancichinetti, S. Fortunato, and J. Kertsz, "Detecting the overlapping and hierarchical community structure of complex networks," *New J. Phys.*, vol. 11, no. 3, pp. 19–44, 2012.
[64] J. Derrac, S. Garca, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 3–18, 2011.
[65] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 1–35, 2011.
[66] L. While, P. Hingston, L. Barone, and S. Huband, "A faster algorithm for calculating hypervolume," *IEEE Trans. Evol. Comput.*, vol. 10, no. 1, pp. 29–38, Feb. 2006.
[67] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," *Proc. Nat. Acad. Sci.*, vol. 104, no. 1, pp. 36–41, 2007.

**Ye Tian** received the B.Sc. degree in software engineering, M.Sc. degree in technology of computer application, and Ph.D. degree in technology of computer application from Anhui University, Hefei, China, in 2012, 2015, and 2018, respectively.

He is currently an Associate Professor with the Institutes of Physical Science and Information Technology, Anhui University. His current research interests include multiobjective optimization methods and their application.

Dr. Tian is the recipient of the 2018 IEEE Transactions on Evolutionary Computation Outstanding Paper Award and the 2020 IEEE Computational Intelligence Magazine Outstanding Paper Award.

**Shangshang Yang** received the B.Sc. degree in network engineering in 2017, from Anhui University, Hefei, China, where he is currently working toward the Ph.D. degree in technology of computer application with the School of Computer Science and Technology.

His current research interests include multiobjective optimization, community detection, deep learning, and data mining.

**Xingyi Zhang** (SM'18) received the B.Sc. degree in mathematics from Fuyang Normal College, Fuyang, China, in 2003, and the M.Sc. degree in applied mathematics and Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology, Wuhan, China, in 2006 and 2009, respectively.

He is currently a Professor with the School of Computer Science and Technology, Anhui University, Hefei, China. His current research interests include unconventional models and algorithms of computation, evolutionary multiobjective optimization, and data mining.

Dr. Zhang is the recipients of the 2018 IEEE Transactions on Evolutionary Computation Outstanding Paper Award and the 2020 IEEE Computational Intelligence Magazine Outstanding Paper Award.