A Network Reduction-Based Multiobjective Evolutionary Algorithm for Community Detection in Large-Scale Complex Networks

Xingyi Zhang, Kefei Zhou, Hebin Pan, Lei Zhang, Xiangxiang Zeng, and Yaochu Jin, *Fellow, IEEE*

***Abstract*—Evolutionary algorithms have been demonstrated to be very competitive in the community detection for complex networks. They, however, show poor scalability to large-scale networks due to the exponential increase of search space. In this paper, we suggest a network reduction-based multiobjective evolutionary algorithm for community detection in large-scale networks, where the size of the networks is recursively reduced as the evolution proceeds. In each reduction of the network, the local communities found by the elite individuals in the population are identified as nodes of the reduced network for further evolution, thereby considerably reducing the search space. A local community repairing strategy is also suggested to correct the misidentified nodes after each network reduction during the evolution. Experimental results on synthetic and real-world networks demonstrate the superiority of the proposed algorithm over several state-of-the-art community detection algorithms for large-scale networks, in terms of both computational efficiency and detection performance.**

***Index Terms*—Community detection, complex network, evolutionary algorithm, large-scale network, multiobjective optimization.**

I. Introduction

MULTIOBJECTIVE evolutionary algorithms (MOEAs) have attracted increasing attention in complex networks for community detection [1]. Generally speaking, community detection is to divide a network into several groups of nodes based on the topology structure of the network, such that nodes in the same group are densely connected and the nodes in different groups are sparsely connected. Community detection is a very important tool for uncovering the information hidden in the complex networks, such as biological networks [2] and social networks [3], where most of these networks usually hold the scale free property, that is, a network whose degree distribution follows a power law [4], [5]. Different from the community detection, clustering is to divide a set of objects into several subsets based on the similarity between the objects, such that the objects in the same subset are more similar than those in different subsets. In the calculation of similarity in clustering, objects are often represented as several key features obtained by using some feature extraction technologies. In this sense, community detection is a special case of clustering and existing clustering methods can be used for community detection, in case that suitable features for each node can be extracted based on the topology structure of the network. There exists some work focusing on extracting features of nodes in networks and clustering methods for community detection, such as [6] and [7].

The first MOEA for community detection in complex networks, called MOGA-net, was suggested by Pizzuti in 2009 [8], where two objectives, community score, and community fitness, were introduced for optimization. Since then, a considerable number of MOEAs have been reported for community detection in complex networks [9]–[15]. For example, Amiri *et al.* [9] employed the harmony search algorithm for community detection in complex networks by optimizing two objectives: 1) maximization of internal links and 2) minimization of external links. Shi *et al.* [11] developed a multiobjective community detection algorithm (MOCD) based on the improved version of Pareto envelope-based selection algorithm PESA-II [16]. Gong *et al.* [13] suggested a decomposition-based multiobjective discrete particle swarm optimization algorithm (MODPSO) to detect communities in complex networks. Recently, Chen *et al.* [15] proposed a teaching-learning-based multiobjective discrete algorithm (MODTLBO/D) for complex networks community detection.

Compared with single-objective EAs, there are two main reasons for adopting an MOEA for community detection in complex networks. First, the use of multiple objectives can overcome some potential disadvantages in optimizing a single objective, such as limited resolution of modularity. As reported in [11] and [13], communities smaller than a certain

Manuscript received February 9, 2018; revised August 20, 2018; accepted September 17, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61822301, Grant 61672033, Grant 61502001, Grant 61876184, and Grant 61502004, in part by the Anhui Provincial Natural Science Foundation for Distinguished Young Scholars under Grant 1808085J06, in part by the Joint Research Fund for Overseas Chinese, Hong Kong and Macao Scholars of the National Natural Science Foundation of China under Grant 61428302, and in part by the U.K. EPSRC under Grant EP/M017869/1. This paper was recommended by Associate Editor S. Mostaghim. *(Corresponding author: Lei Zhang.)*

X. Zhang, K. Zhou, H. Pan, and L. 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 230039, China (e-mail: xyzhanghust@gmail.com; kefei_zhou@163.com; bimk_phb@163.com; zl@ahu.edu.cn).

X. Zeng is with the Department of Computer Science, Xiamen University, Xiamen 361005, China (e-mail: xzeng@xmu.edu.cn).

Y. Jin is with the Department of Computer Science, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: yaochu.jin@surrey.ac.uk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

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

Digital Object Identifier 10.1109/TCYB.2018.2871673

size cannot be identified by only optimizing modularity and multiobjective optimization is an effective method to break through the modularity limitation. Second, MOEAs can return a set of Pareto optimal solutions instead of a single optimal solution. Each of these solutions corresponds to a different tradeoff between the multiple objectives and thus a set of network divisions at different hierarchical levels can be easily obtained by MOEAs as pointed out in [10]. This means that MOEAs can provide a promising method for hierarchical community structure detection, which is a very hot topic in community detection [17], [18].

MOEAs have demonstrated the superiority in detecting communities for complex networks, and many competitive MOEA-based community detection algorithms have been developed. However, these MOEAs still suffer from poor scalability to large-scale networks due to the curse of dimensionality. The main reason is due to the fact that the individual length of encoding a network is proportional to the number of nodes in the network. This means that the community detection in large-scale networks becomes a multiobjective optimization problem (MOP) with large scale, which is much more challenging than MOPs with a relatively small number of decision variables despite the fact that some specially tailored strategies have been proposed in EAs [19], [20].

To address this issue, in this paper, we propose a network reduction-based MOEA (RMOEA), for community detection in large-scale complex networks. In the proposed RMOEA, the size of networks is recursively reduced as the evolution proceeds, thus providing an effective method to deal with large-scale networks. Specifically, the main contributions of this paper can be summarized as follows.

1) A network reduction strategy is suggested for community detection in large-scale complex networks based on the local communities found by the elite individuals in the evolution. With this strategy, several local communities of a network are identified as nodes of the reduced network. This means that the size of the networks being handled is reduced, and thus the scalability of MOEAs to large-scale networks can be considerably enhanced.
2) Based on the proposed reduction method, an MOEA named RMOEA is suggested for community detection in large-scale complex networks. In RMOEA, the two metrics KKM and RC for measuring intralink and interlink densities [13] are adopted as two objectives for optimization. In addition, a local community repairing strategy is suggested in RMOEA to correct some misidentified nodes in the found local communities after each reduction of the network.
3) The effectiveness and efficiency of the proposed RMOEA are verified both on synthetic benchmark and seven real-world networks. Experimental results demonstrate that the proposed RMOEA performs better than the state-of-the-art EA-based and non-EA-based community detection algorithms on large-scale complex networks.

The rest of this paper is organized as follows. The definition of MOCD and related work on MOEAs for community detection are introduced in Section II. The proposed MOEA is presented in Section III, with a detailed description of the network reduction strategy. In Section IV, we empirically verify the effectiveness and efficiency of the proposed algorithm by comparing it with several state-of-the-art community detection algorithms. Finally, the conclusion and some future remarks are given in Section V.

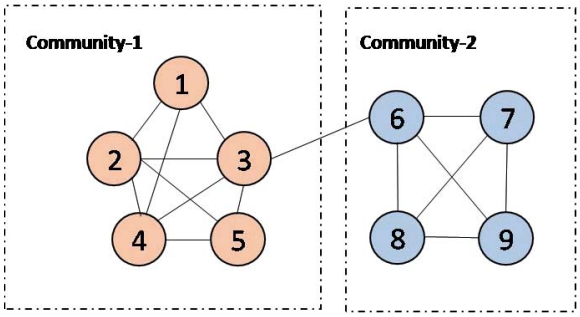


Fig. 1. Complex network having nine nodes, where the network is divided into two communities, one consisting of nodes 1–5 and the other consisting of nodes 6–9.

II. Problem Formulation and Related Work

In this section, we first present the MOCD, and then recall the related work on MOEA-based community detection algorithms.

A. Multiobjective Community Detection

Community detection is a procedure where nodes in the network are divided into different groups based on the density of links between nodes. The links between nodes in the same group (called intralink density) need to be as dense as possible, whereas links between nodes in different groups (called interlink density) should be sparse enough. Such groups of nodes are often called communities in complex networks. Fig. 1 provides an illustrative example of a complex network with nine nodes. From the figure, it can easily be observed that the network is divided into two communities, one consisting of five nodes $\{1, 2, 3, 4, 5\}$, and the other consisting of four nodes $\{6, 7, 8, 9\}$. In the following, we only consider the detection of communities similar to those shown in Fig. 1, where no node is shared by two or more communities. This means that we here do not consider overlapping community detection.

To perform community detection, several criteria have been proposed to measure intralink and interlink densities [1]. In this paper, we formulate the community detection as a bi-objective optimization problem by minimizing the kernel k-means (KKM) and the ratio cut (RC) which has been popularly adopted in recent works [13], [21], [22]. Given a network $G = (V, E)$ ($|V| = n$, $|E| = m$), and A is the adjacent matrix of G, the bi-objective minimization problem can be formally defined as follows:

$$\begin{aligned} \texttt{minimize}\ \ & F(\mathcal{C}) = (\mathrm{KKM}(\mathcal{C}), \mathrm{RC}(\mathcal{C})) \\ & \mathrm{KKM}(\mathcal{C}) = 2(n-k) - \sum_{i=1}^{k} \frac{L(V_i, V_i)}{|V_i|} \\ & \mathrm{RC}(\mathcal{C}) = \sum_{i=1}^{k} \frac{L(V_i, \overline{V_i})}{|V_i|} \end{aligned} \tag{1}$$

where $\mathcal{C} = \{V_1, V_2, \ldots, V_k\}$ is a division of G with k communities $V_1, V_2, \ldots, V_k$, $L(V_i, V_i) = \sum_{i \in V_i, j \in V_i} A_{ij}$ and $L(V_i, \overline{V_i}) = \sum_{i \in V_i, j \in \overline{V_i}} A_{ij}$ denote the internal and external link densities of nodes in V_i.

As reported in [13], the reason for adopting the above two objectives is due to the fact that the KKM is a decreasing function of the number of communities whereas the opposite trend happens to the RC, which means that KKM and RC are conflicting objectives. As indicated in the definitions, the KKM is to measure the sum of the link density of intracommunities and the RC measures the link density of intercommunities. The minimization of both KKM and RC can ensure that the links within a community are dense whereas the links between the communities are spare, which is coincidence with the requirements of community detection.

B. Related Work on MOEA-Based Community Detection Algorithms

We here review a few MOEAs widely used in community detection for complex networks. It is worth noting that there also exist some competitive single-objective EA-based community detection algorithms, such as community detection algorithm-based single-objective genetic algorithm (GA-net) [23], genetic algorithm for community detection [24], chemical reaction optimization with dual-representation for community detection [25], and cooperative co-evolutionary module identification for cancer disease community discovery (CoCoMi) [26].

The idea of formulating community detection as an MOP was first proposed by Pizzuti in 2009 [8], where two objectives, community score and community fitness, were considered. For solving the MOP, a multiobjective algorithm, termed MOGA-Net, was also suggested by Pizzuti in [8] based on the fast nondominated sorting genetic algorithm (NSGA-II) [27], with a detailed empirical evaluation presented in [10]. In MOGA-Net, MOEAs demonstrated two interesting superiorities for community detection in complex networks in comparison with single-objective EAs. On the one hand, the detection performance can be considerably enhanced by optimizing multiple conflicting objectives simultaneously, which can overcome some drawbacks existing in single-objective optimization, such as the resolution limitation in modularity Q [28]. On the other hand, MOEAs can provide a set of hierarchial community structures, since each of the Pareto optimal solutions corresponds to a different tradeoff between the objectives being optimized.

In recognizing the superiorities of MOEAs in community detection, Shi *et al.* [11] proposed an MOCD algorithm for complex networks based on the well-known MOEA and PESA-II [16]. In [11], it was shown that the communities detected by MOCD are more accurate than those found by well-established single-objective community detection algorithms on both synthetic and real-world networks. Gong *et al.* [12] proposed a community detection algorithm, MOEA/D-Net, based on the MOEA with decomposition, where two objectives, negative ratio association and ratio cut, were adopted for optimization. Experimental results demonstrated the competitiveness of MOEA/D-Net on small-scale networks, such as the extension of GN benchmark network which consists of 128 nodes divided into four communities of 32 nodes each [12].

To further enhance the performance of MOEA-based community detection algorithms, Gong *et al.* [13] developed an MODPSO for community detection in complex networks, where two metrics, KKM and RC, were adopted as two objectives. In MODPSO, a problem-specific population initialization method was also suggested, on the basis of the label propagation which has been widely adopted for designing community detection algorithms in complex networks. It was shown that MODPSO outperformed existing single-objective EAs and MOEAs for community detection in complex networks with more nodes, e.g., LFR benchmark network with 1000 nodes and power grid with 4941 nodes [13].

There are a few other MOEA-based community detection algorithms which were developed based on different swarm intelligence algorithms, such as the modified harmony search algorithm with a chaotic local search for community detection [29], multiobjective immune algorithm-based community detection algorithm [30], community detection algorithm based on an enhanced firefly algorithm [31], multiobjective discrete backtracking search optimization algorithm with decomposition for community detection [32], and MODTLBO/D [15]. There are also some interesting work focused on overlapping community detection [21], [22], [33] or other kinds of networks, such as signed networks [33]–[35] and dynamic networks [36], [37].

The MOEA-based community detection algorithms mentioned above have demonstrated the competitiveness in a variety of networks, their performance, however, will considerably deteriorate on large-scale complex networks due to the curse of dimensionality. For all existing community detection algorithms based on MOEAs, the length of individuals used for encoding a network is closely related to the number of nodes in the network, and this length remains unchanged during the whole process of evolution. Hence, the search space will exponentially increase as the number of nodes in networks increases, leading to the ineffectiveness of MOEA-based community detection algorithms on large-scale networks.

To address this issue, in this paper, we propose a network RMOEA for community detection in large-scale networks. In RMOEA, the size of the network to be handled is recursively reduced in evolution by identifying the local communities found by elite individuals as nodes of the reduced networks for further evolution. In other words, the length of individuals for encoding a network in the proposed RMOEA will recursively decrease in evolution, thereby considerably reducing the search space and enhancing the community detection performance in large-scale networks. In Section IV, we will verify the competitive performance of RMOEA on large-scale complex networks in terms of both computational efficiency and detection performance, in comparison with several state-of-the-art community detection algorithms.

Algorithm 1 EliteSelection(G, P)

Input: G: complex network; P: population;
Output: P': set of elite individuals;

```
pop ← the number of individuals in P;
for i = 1 to pop do
    Q(i) ← calculate the Q value of division corresponding to individual p_i;
    C_num(i) ← calculate the number of communities in the division corresponding to individual p_i;
end for
P' ← obtain the non-dominated solutions in P by using Q and C_num for elite individual selection;
```

Algorithm 2 Evo-Reduction(G, P)

Input: G: complex network; P: population;
Output: G_R: the reduced network; P_R: population for the reduced network;

```
G_R ← G;
P_R ← P;
P' ← EliteSelection (G_R, P);
ind ← find the individual in P' with the largest number of communities;
l ← the number of communities in ind;
C ← the communities C_1, ..., C_l corresponding to ind;
for i = 1 to l do
    C_i^s ← C_i;
    for each ind' ∈ P' and ind' ≠ ind do
        C' ← the communities C'_1, ..., C'_t corresponding to ind';
        C'_j ← find one community C'_j, j ∈ {1, ..., t}, satisfying that C_i ∩ C'_j has the most number of nodes;
        C_i^s ← C_i^s ∩ C'_j;
    end for
    G_R ← merge C_i^s into one node in G_R;
    P_R ← merge the genes of nodes of C_i^s into one gene for each individual in P_R;
end for
```

III. Proposed Algorithm RMOEA

In this section, we present the details of the proposed RMOEA for community detection in large-scale networks. First, we present the network reduction method developed in RMOEA, which is the key component of RMOEA, and then give a local community repairing strategy suggested in RMOEA for correcting misidentified nodes in the reduced network. Finally, the general framework of the proposed RMOEA is presented.

A. Network Reduction Method

To perform community detection in large-scale networks, a network reduction method is suggested in the proposed RMOEA to recursively reduce the size of the networks in evolution. The main idea of reduction method is motivated by the following observation. In complex networks, there often exist some nodes which are easy to be checked that they belong to the same community, thus will be found by MOEAs after a small number of generations. The proposed network reduction method utilizes the local community information found in the evolution and considers each of the found local communities as a whole in further evolution. In other words, the nodes in the found local community will keep unchanged at later stage of evolution.

Fig. 2 presents the main idea of the network reduction method, where a network G with 7 nodes is considered and we assume that two individuals *ind*1 and *ind*2 have been found by an MOEA for community detection in the network. As shown in Fig. 2(c), the individual *ind*1 corresponds to a division consisting of four communities $\{1, 2, 3, 4\}$, $\{5\}$, $\{6\}$, and $\{7\}$, and *ind*2 corresponds to a division of five communities $\{1, 2, 3\}$, $\{4\}$, $\{5\}$, $\{6\}$, and $\{7\}$, according to the locus-based adjacency encoding scheme suggested in [38]. It can be seen that both individuals *ind*1 and *ind*2 identify nodes 1–3 as in one community. Hence, we consider the local community $\{1, 2, 3\}$ as a whole in further evolution, which is regarded as a node in the reduced network depicted in Fig. 2(d). In other words, in further evolution, we will not check the possibility that nodes 1–3 do not belong to the same community. By using the local community found by individuals *ind*1 and *ind*2, the network having seven nodes is reduced to a network with five nodes, thus considerably enhancing the community detection performance in large-scale networks.

Due to the fact that not all individuals in population have a good local information in evolution, some elite individuals from the population are selected for determining the local communities more correctly. To this end, we adopt two metrics, modularity Q [39] and the number of communities *C_num*, to select elite individuals from the population. The modularity Q is a widely used metric in complex networks for evaluating the quality of obtained communities. The larger the value of Q is, the better quality the found communities have. The reason for choosing the number of communities *C_num* as a metric lies in the fact that we would like to obtain the local communities as small as possible, since these communities will be regarded as nodes in the reduced network. For a given network, a large value of *C_num* implies that each local community has a small size, which diminishes the possibility of incorrectly identified local communities. The effectiveness of Q and *C_num*-based selection in the proposed RMOEA can be found in the supplementary material I. The elite individual selection is performed as follows. For a population P, we calculate the values of Q and *C_num* for each individual in P. The individuals which are not dominated by any individual in P according to Q and *C_num* are selected as elite individuals to determine the local communities. The main procedure of elite individual selection is presented in Algorithm 1.

Algorithm 2 presents the procedure of the proposed network reduction method based on the local communities found by elite individuals in evolution. Suppose k individuals in population P are selected as elite individuals for determining the local communities, and the set of elite individuals is denoted as P'. Note that each individual represents a network division

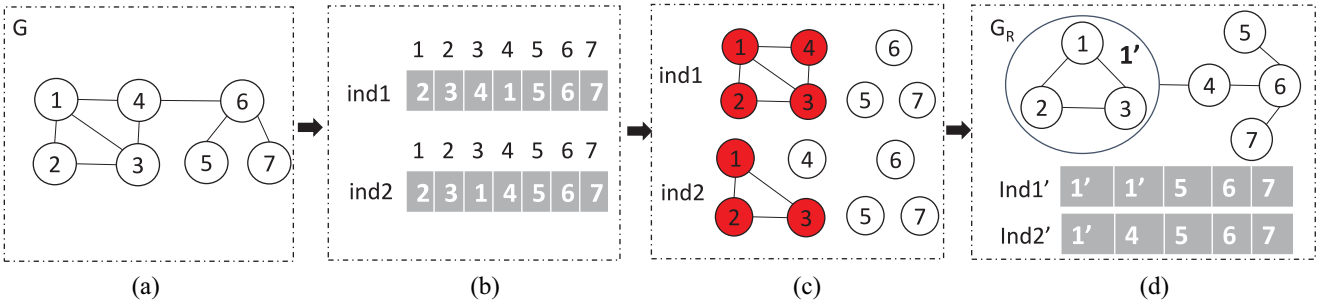


Fig. 2. Example to illustrate the main idea of the proposed network reduction method in the evolution of population. (a) Complex network G. (b) Two individuals *ind*1 and *ind*2 having been found for network G, where locus-based individual representation scheme is adopted [10]. (c) Two divisions corresponding to individuals *ind*1 and *ind*2. The division of *ind*1 consists of four communities $\{1, 2, 3, 4\}$, $\{5\}$, $\{6\}$, and $\{7\}$, and the division of *ind*2 is $\{1, 2, 3\}$, $\{4\}$, $\{5\}$, $\{6\}$, and $\{7\}$. (d) Reduced network G_R based on the local communities found by individuals *ind*1 and *ind*2, where the local community $\{1, 2, 3\}$ is identified as a node in the reduced network since both individuals recognize nodes 1, 2, and 3 as in one community. *ind*1′ and *ind*2′ are the new individuals of *ind*1 and *ind*2 for the reduced network.

consisting of several communities. The proposed network reduction method is performed as follows. First, the individual *ind* with the largest number of communities in P' is selected. Second, the communities corresponding to *ind* are used to obtain local communities based on whether all the nodes in the communities of *ind* are identified to belong to one community by the remaining elite individuals. To be specific, for each community C_i in *ind*, if all nodes in C_i are identified as in one community by the remaining $k-1$ elite individuals, then C_i is considered as a local community. Otherwise, we consider the maximal subset of C_i which are identified in one community by the remaining $k-1$ elite individuals as a local community. This above operation halts until all communities in *ind* are checked. Lastly, a reduced network G_R is obtained by merging each local community into one node in G, and a population P_R for the reduced network G_R is obtained by merging the genes of nodes of each local community into one gene for all individuals in P. For each individual, an index of the local community is assigned to the merged gene and the community information on the rest genes in the individual keeps unchanged.

Take the complex network depicted in Fig. 2 as an illustrative example to show the procedure of the proposed network reduction method given by Algorithm 2. Let us assume that *ind*1 and *ind*2 are the elite individuals selected from population P for determining the local communities, as shown in Fig. 2(b). Since the individual *ind*2 contains more communities than *ind*1 (4 communities are contained in *ind*1, namely, $\{1, 2, 3, 4\}$, $\{5\}$, $\{6\}$, and $\{7\}$, and 5 communities are contained in *ind*2, namely, $\{1, 2, 3\}$, $\{4\}$, $\{5\}$, $\{6\}$, and $\{7\}$), the 5 communities contained in *ind*2 are used to obtain the local communities. For community $\{1, 2, 3\}$, all nodes 1–3 in the community are identified as in one community by individual *ind*1, and thus $\{1, 2, 3\}$ is considered as a local community. As for community $\{4\}$, it is considered as a local community due to the fact that only one node is contained in the community. This is also the case for communities $\{5\}$, $\{6\}$, and $\{7\}$. In this way, five local communities $\{1, 2, 3\}$, $\{4\}$, $\{5\}$, $\{6\}$, and $\{7\}$ are obtained for the network and thus the network is reduced to a network with five nodes {1', 4, 5, 6, 7}, where the local community $\{1, 2, 3\}$ is merged into one node 1'. For the reduced network, individuals *ind*1 and *ind*2 are changed to *ind*1′ and *ind*2′ by assigning 1' to the merged gene and keeping the information on the rest genes unchanged.

The above network reduction is based on the local communities found by elite individuals in evolution. There are also some local communities which can easily be determined by checking the local topology structure of networks. To enhance the scalability of the proposed RMOEA in large-scale networks, we suggest a prereduction method based on the local topology structure of networks, which is performed before the evolution starts.

Algorithm 3 presents the procedure of prereduction by finding local communities before the evolution. The algorithm is performed as follows. First, a node s is randomly selected from the network to be detected, and a subgraph G^{LS} consisting of nodes s^c (a neighboring node of s which has the largest degree) and s^{cc} (a neighboring node of s^c which shares the largest number of neighbors with s^c), as well as all common neighbors of s^c and s^{cc}, is obtained. Second, the subgraph G^{LS} is extended by adding each neighbor s' of nodes in G^{LS} whose number of links with G^{LS} is larger than one half of degree of s', and the extended subgraph G^{LS} is regarded as a local community of the network. After a local community is found, the algorithm starts to find another local community by using the above operations from the remaining nodes of the network. This process repeats until all nodes in the network have been assigned to local communities. It is worth mentioning that the random choice of nodes (step 4 in Algorithm 3) has little bias on the final obtained local communities, and the analysis can be found in the supplementary material II.

Fig. 3 presents an example to illustrate the procedure of prereduction performed before the evolution, where a network having seven nodes is considered. Assume that node 1 is randomly selected for finding a local community. The subgraph $\{1, 2, 3, 5\}$ depicted in Fig. 3(b) is obtained by using the above operations starting from node 1, where node 3 is the neighbor of node 1 which has the largest degree and for all neighbors of node 3, nodes 1, 2, and 5 are those which have the largest number of common neighbors with node 3. As shown in Fig. 3(c), node 4 can be added into the subgraph $\{1, 2, 3, 5\}$, since all links associated with node 4 are connected to nodes

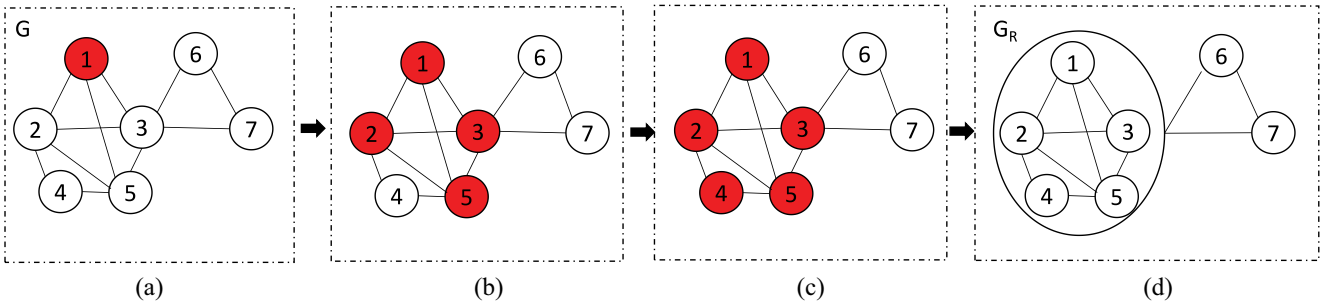


Fig. 3. Illustrative example of the prereduction. (a) Complex network G consisting of seven nodes (assume that node 1 is randomly selected for finding a local community). (b) Subgraph determined by node 1, consisting of four nodes 1, 2, 3, and 5. (c) Local community determined by node 1 by adding node 4 into the subgraph. (d) Reduced network G_R based on the found local community, where the local community $\{1, 2, 3, 4, 5\}$ is regarded as a node in the reduced network.

Algorithm 3 Prereduction(G)

Input: G: complex network;
Output: G_R: the reduced network;

$G_R \leftarrow G$;
Nodes $\leftarrow$ nodes in G;
while *Nodes* $\neq \emptyset$ **do**
  $s \leftarrow$ randomly select one node from *Nodes*;
  $s^c \leftarrow$ find the node in the neighborhood of s which has the largest degree;
  $s^{cc} \leftarrow$ find the node in the neighborhood of s^c which has the largest number of common neighboring nodes with s^c;
  $G^{LS} \leftarrow$ get the subgraph of s;
  $NB \leftarrow$ find the neighbors which are not in G^{LS}, for each node in G^{LS};
  $NC \leftarrow \emptyset$;
  while $NB \neq NC$ **do**
    $NC \leftarrow NB$;
    $t \leftarrow$ randomly select one node from NB;
    if the number of links between t and G^{LS} is larger than one half of degree of t **then**
      $G^{LS} \leftarrow G^{LS} \cup \{t\}$;
      $NB \leftarrow$ find the neighbors which are not in G^{LS}, for each node in G^{LS};
    end if
  end while
  $G_R \leftarrow$ merge all nodes in G^{LS} into one node in G_R;
  Nodes $\leftarrow$ remove nodes in G^{LS} from *Nodes*;
end while

in subgraph $\{1, 2, 3, 5\}$. The network is reduced to a network with 3 nodes by merging the local community $\{1, 2, 3, 4, 5\}$ into one node, hence the size of network to be detected is considerably reduced before the evolution.

B. Local Community Repairing Strategy

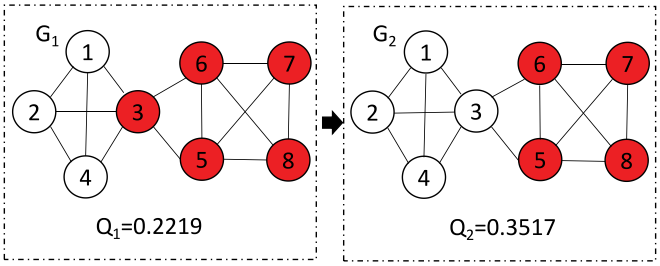
In the proposed network reduction method, the local communities may be incorrectly found by elite individuals in evolution, especially at the early stage of search. In other words, the proposed network reduction method may find a local community whose nodes do not belong to one community, which easily leads to the performance deterioration of RMOEA. To this end, we suggest a local community repairing strategy to correct the misidentified nodes in the reduced network.

Algorithm 4 presents the main procedure of the local community repairing strategy, which is performed as follows. Assume that G_R is a reduced network and P_R is the population associated with G_R. First, the set of elite individuals P' is selected from population P_R by Algorithm 1. Second, the local communities found by each elite individual are checked whether they contain some nodes which do not belong to the same community with the rest of nodes in the local communities. To be specific, for each local community found by elite individuals, all nodes in the local community will be checked whether there exists at least one neighbor not in the local community. A node is regarded as misidentified one if a better Q value of communities is obtained by moving the node to the communities of its neighbors. Lastly, each misidentified node is removed from the merged nodes in the reduced network G_R as a separate node, and a corresponding gene is added for this node in each individual in P_R, where the value of the gene is randomly assigned to an index of a neighbor of the node.

Fig. 4 presents an illustrative example of the local community repairing strategy, where we assume that a local community $\{3, 5, 6, 7, 8\}$ has been found by elite individuals in the population, as shown in Fig. 4(a). The local community repairing strategy first finds node 3 in the local community $\{3, 5, 6, 7, 8\}$, which is a node with at least one neighbor not in the local community. Then, the strategy checks whether the quality of communities can be improved by moving node 3 to another community $\{1, 2, 4\}$. Since the Q value increases after the movement, node 3 needs to be removed from the local community $\{3, 5, 6, 7, 8\}$. In this way, the local community $\{3, 5, 6, 7, 8\}$ is repaired, by which the quality of local communities found by elite individuals in evolution can be greatly improved, leading to fewer misidentified nodes in the reduced network.

C. General Framework of RMOEA

The proposed RMOEA adopts a similar framework with MOEA/D-Net [12], which is presented in Algorithm 5. Let *pop* be the population size, then the weight vectors $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{pop}\}$ are a set of weight vectors evenly spread



(a) (b)

Fig. 4. Example to show the main idea of the local community repairing strategy. (a) Local community $\{3, 5, 6, 7, 8\}$ found by using elite individuals in evolution. (b) Repaired local community $\{5, 6, 7, 8\}$ by moving node 3 to the community $\{1, 2, 4\}$ since the quality of communities, Q, increases after the movement.

Algorithm 4 LocalCommunityRepairing(G_R, P_R)

Input: G_R: the reduced network; P_R: population for the reduced network;
Output: G_R^r: the repaired network; P_R^r: population for the repaired network;

$P' \leftarrow$ EliteSelection(G_R, P_R);
for $i = 1$ to $|P'|$ **do**
  $k \leftarrow$ the number of communities corresponding to p_i' in P';
  $\mathcal{C} \leftarrow$ the communities $C_1, C_2, \ldots, C_k$ corresponding to p_i';
  $Q_i' \leftarrow Modularity(p_i', G_R)$;
  for $j = 1$ to k **do**
    $Nodes \leftarrow$ select nodes in C_j if there exist one neighbor not in C_j;
    while $Nodes \neq \emptyset$ **do**
      $s \leftarrow$ randomly select one node from $Nodes$;
      $p_i'' \leftarrow$ move s to the communities of its neighbors;
      $Q_i'' \leftarrow Modularity(p_i'', G_R)$;
      if $Q_i'' > Q_i'$ **then**
        $Fault\_Nodes \leftarrow Fault\_Nodes \cup \{s\}$;
      end if
      $Nodes \leftarrow$ remove s from $Nodes$;
    end while
  end for
end for
$G_R^r \leftarrow$ remove nodes in $Fault\_Nodes$ from merged nodes in G_R as a separate node;
$P_R^r \leftarrow$ add the genes of nodes of $Fault\_Nodes$ for each individual in P_R;

on $\lambda_i^1 + \lambda_i^2 = 1$, where $\lambda_i = \langle \lambda_i^1, \lambda_i^2 \rangle$, $\lambda_i^1, \lambda_i^2 \in [0, 1]$ and $1 \leq i \leq pop$. The weight vectors λ_i, $1 \leq i \leq pop$, are used to decompose the MOCD problem into pop single-objective subproblems according to the following formula:

$$\text{minimize } g^{te}(x|\lambda_i, z^*) = \max_{j=1}^{2} \left\{ \lambda_i^j \cdot \left(|F_j(x) - z^*| \right) \right\}$$

where $z^* = \langle z_1^*, z_2^* \rangle$ is the reference point in which each z_j^*, $1 \leq j \leq 2$, is the minimal value on the jth objective in population.

Algorithm 5 General Framework of RMOEA

Input: G: complex network, $maxgen$: maximum number of generations, pop: population size, λ: weight vectors $\{\lambda_1, \lambda_2, \ldots, \lambda_{pop}\}$, ns: the size of neighbourhood, p_c: crossover probability, p_m: mutation probability, T: the number of reductions in evolution;
Output: The final population;

$G_R \leftarrow$ Pre-Reduction (G);
Initialize the population $P = \{p_1, p_2, \ldots, p_{pop}\}$;
Initialize reference point z^*;
for $i = 1$ to pop **do**
  $N_i \leftarrow$ find the ns individuals from P with the nearest Euclidean distance to the weight vector λ_i;
end for
$P_R \leftarrow P$;
for $i = 1$ to $maxgen$ **do**
  if $i \mid (\lceil (maxgen + 1)/(T + 1) \rceil) == 0$ **then**
    $[G_R, P_R] \leftarrow$Evo-Reduction(G_R, P_R);
    $[G_R, P_R] \leftarrow$LocalCommunityRepairing(G_R, P_R);
    Update P_R and reference point z^* based on G_R;
  end if
  for $j = 1$ to pop **do**
    Randomly select one individual from N_j;
    Generated children by crossover and mutation operators;
    Update individuals in N_j;
    Update reference point z^*;
  end for
end for

The RMOEA consists of the following four steps. At the first step, the size of the network to be detected is reduced by using the local topology structure of the network according to the proposed prereduction method in Algorithm 3. At the second step, a population with pop individuals is initialized based on the locus-based adjacency encoding schema, which has been widely used in EA-based community detection algorithms [10]–[12], [23]. Specifically, each gene of individuals in P is first generated by randomly assigning the index of neighbors of the node associated with this gene or the index of the node itself, and then the index of each node in individuals is recursively replaced by the index that most neighbors of this node shares until the indexes of nodes are not changed. The reference point z^* is initialized by using the best values of KKM and RC in the initial population. For each weight vector λ_i, $1 \leq i \leq pop$, the Euclidean distances from all individuals in population P to weight vector λ_i are calculated and ns individuals in P with the nearest Euclidean distances to λ_i are regarded as the neighbors of λ_i, denoted as N_i, where ns is a predefined parameter.

At the third step, the size of network is further reduced by using elite individuals according to the proposed network reduction method in Algorithm 2 for each ith generation satisfying that $i \mid (\lceil \text{maxgen} + 1)/(T + 1) \rceil) == 0$, where the symbol "|" is the remainder of the division between i and the expression $\lceil (\text{maxgen} + 1)/(T + 1) \rceil$, and T is a

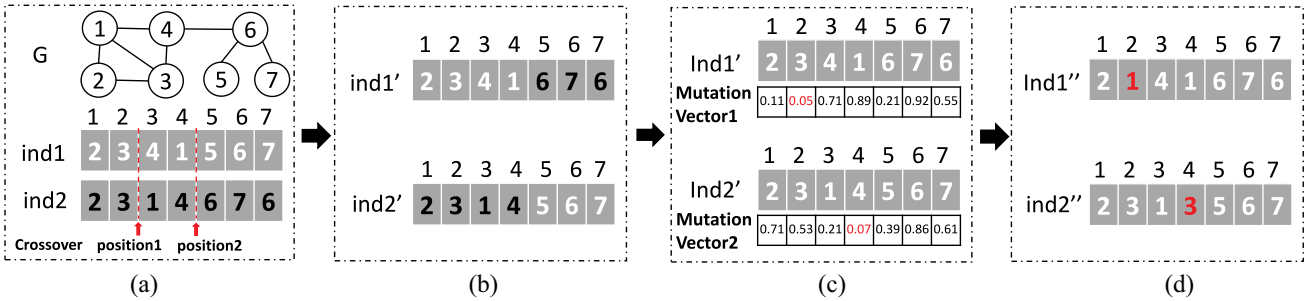


Fig. 5. Illustrative example to illustrate the main process of crossover and mutation. (a) Complex network G and two individuals $ind1$ and $ind2$ with crossover $position1$ and $position2$ generated randomly. For each crossover process, a randomly number $rand$ is generated from $\{1, 2, 3\}$. rand $= 1, 2, 3$ mean that the parts before $position1$, between $position1$ and $position2$, and after $position2$ of $ind1$ and $ind2$ are mutually exchanged, respectively. (b) New $ind1'$ and $ind2'$ are generated by crossover operation for rand $= 3$. (c) Two mutation vectors for $ind1'$ and $ind2'$ are generated with values randomly from [0, 1], if the value of the node in mutation vector is smaller than 0.1, the corresponding position of individual will mutate to an index in its neighboring nodes. (d) New individuals $ind1''$ and $ind2''$ are generated by mutation operation.

parameter for controlling the number of reductions in evolution. The suggested local community repairing strategy in Algorithm 4 is conducted to correct the misidentified nodes in the reduced network after each reduction. At the last step, for each individual p_j, $1 \leq j \leq pop$, in P, one individual ind is randomly selected from N_j. The two individuals p_j and ind are used to generate the child chd by crossover and mutation operators adopted in [12]. Fig. 5 gives an illustrative example of crossover and mutation operations. If the Tchebycheff value of chd is better than an individual ind' in N_j, then replace ind' with chd and update reference point z^*. The RMOEA keeps going until the maximum number of generations is reached. The complexity analysis of RMOEA can be found in the supplementary material III.

IV. Experimental Results and Analysis

In this section, we empirically verify the performance of the proposed RMOEA in terms of both computational efficiency and detection performance. Specifically, the experiments are designed as follows.

A. Experimental Design

1) Comparison Algorithms: The performance of RMOEA is compared with five popular EA-based community detection algorithms for complex networks, namely, CoCoMi [26], MOCD [11], MOGA-net [10], MODTLBO/D [15], and MODPSO [13], and two representative non-EA-based community detection algorithms, namely, Louvain [40] and SSCF [7]. Among the five compared EA-based algorithms, CoCoMi is a single-objective EA, and the other four ones belong to MOEAs. The parameters of all compared algorithms are set to the values recommended in [13], which is listed in Table I. In addition, all the algorithms are implemented with the same language, MATLAB. The MATLAB source codes of all compared algorithms are obtained from their authors, with the only three exceptions of MOCD, MODPSO, and Louvain (the source codes of these three algorithms obtained from the authors are written in C++). To make a fair comparison, we have rewritten the codes of MOCD, MODPSO, and Louvain algorithms in MATLAB language and tried our best to optimize the codes to make them work as efficiently as possible. In RMOEA, the number of reductions in evolution is set to $T = 2$ in the experiments.

TABLE I
Parameter Settings of the Compared Algorithms, Where *pop* Denotes the Population Size, *maxgen* Is the Maximum Generations of the Algorithm, *pc* and *pm* Are the Crossover and Mutation Possibility, and *ns* Is the Neighborhood Size

Algorithm	*pop*	*maxgen*	*pc*	*pm*	*ns*	Reference
CoCoMi	100	100	0.9	0.1	–	[26]
MOCD	100	100	0.9	0.1	–	[11]
MOGA-net	100	100	0.9	0.1	–	[10]
MODTLBO/D	100	100	0.9	0.1	40	[15]
MODPSO	100	100	–	0.1	40	[13]
Louvain	–	–	–	–	–	[40]
SSCF	–	–	–	–	–	[7]
RMOEA	100	100	0.9	0.1	40	[ours]

TABLE II
Information of the Seven Real-World Networks Used in the Experiments

Network	Node number	Link number	Average degree
Net-science	1,589	2,742	3.46
blogs	3,982	6,803	3.42
ca-GrQc	5,242	14,496	5.53
ca-HepTh1	9,877	25,998	5.26
ca-HepTh2	12,008	118,521	19.74
ca-AstroPh	18,772	198,080	21.10
ca-CondMat	23,133	93,468	8.08

2) Test Data Sets: The compared algorithms are tested on both synthetic benchmark networks and real-world networks. The synthetic networks we consider are the LFR networks developed by Lancichinetti *et al.* [41], which are the most widely adopted benchmark networks for testing the performance of algorithms in community detection. Compared with other synthetic networks, the LFR networks can reflect some important features of complex real-world systems, since the distributions of node degree and community size in LFR networks are both power laws with tunable exponents. The real-world networks we consider include two real networks *net-science*, *blogs* used in [42] and five real networks *ca-GrQc*, *ca-GrQc*, *ca-Hepth*, *ca-AstroPh*, and *ca-CondMat* developed by the Stanford Network Analysis Project [43]. Table II lists the detailed information of the seven real-world networks.

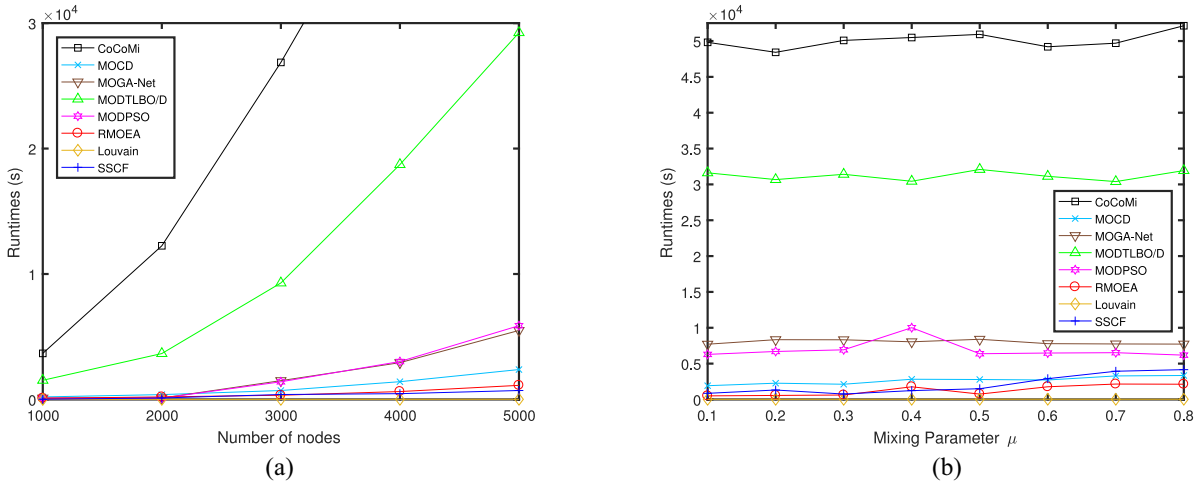


Fig. 6. Runtime (s) of the eight algorithms by averaging over 20 runs on different LFR networks. (a) Runtime on LFR networks with different numbers of nodes. (b) Runtime on LFR networks with different values of μ.

3) Evaluation Metrics: To evaluate the quality of detected communities, two well-known performance metrics, namely, normalized mutual information (NMI) [44] and modularity (Q) [39] are adopted in this paper. For synthetic networks, we use the NMI to measure the similarity between the true community results and the detected ones due to the fact that the ground truth of these networks is known. For real-world networks, the Q is used as the performance metric since their ground truth is not known. Formally, the NMI is defined as

$$\mathrm{NMI}(A,B)=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}C_{ij}\log\left(C_{ij}\dot{n}/C_{i.}C_{.j}\right)}{\sum_{i=1}^{C_A}C_{i.}\log(C_{i.}/n)+\sum_{j=1}^{C_B}C_{.j}\log\left(C_{.j}/n\right)}$$

where C_A (C_B) is the number of communities in division A (B), C is the confusion matrix whose element C_{ij} is the number of nodes shared by community i in division A and by community j in division B, $C_{i.}$ ($C_{.j}$) is the sum of elements of C in row i (column j), and n is the number of nodes of the network. If $A=B$, then $\mathrm{NMI}(A,B)=1$; if A and B are totally different, then $\mathrm{NMI}(A,B)=0$. The modularity Q can be formulated as follows:

$$Q=\sum_{s=1}^{k}\left[\frac{l_s}{M}-\left(\frac{d_s}{2M}\right)^2\right]$$

where M is the number of links of the network, k is the number of detected communities, l_s is the number of links connecting nodes inside the community s, and d_s is the sum of degrees of nodes in s. For both NMI and Q, a larger value indicates a better detection performance.

All simulations reported in this paper are conducted on a small ThinkServer with a Intel XeonM CPU E5-2650, 2.2 GHz, 64 GB memory and the Windows server 2012 R2 standard 64-bit operating system. For each test network, 20 independent runs are performed for each compared algorithm and the mean is recorded. In the following experiments, the solution with the best value of Q in the obtained nondominated solution set is adopted for the final result for all MOEA-based community detection algorithms, since this practice has been widely adopted in the existing MOEA-based community detection algorithm in comparing the performance [13], [21], [22], [33].

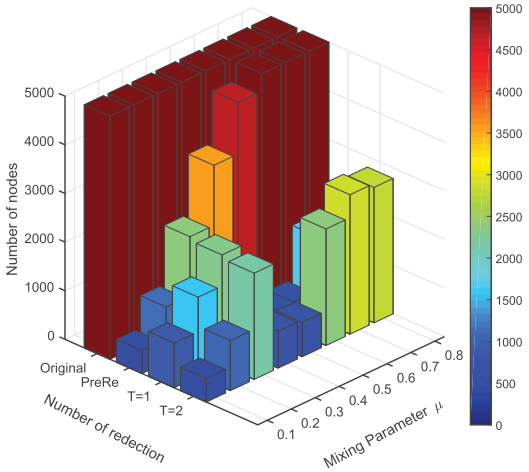


Fig. 7. Number of nodes in the reduced network after the prereduction (*PreRe*), the first ($T=1$) and second ($T=2$) reductions by the proposed RMOEA averaging over 20 runs on LFR networks with different values of μ.

B. Experiments on Synthetic Benchmark Networks

In the experiments, the following three groups of LFR networks are considered. The first group of LFR networks consists of networks whose size ranges from 1000 to 5000 with interval 1000, and the mixing parameter μ is fixed as 0.25. The second group of LFR networks consists of networks whose mixing parameter ranges from 0.1 to 0.8 with interval 0.1, and the size of networks is fixed as $n=5000$. The third group of LFR networks consists of networks whose size ranges from 10 000 to 50 000 with interval 10 000, and the mixing parameter μ is fixed as 0.25. For the three groups of LFR networks, the remaining parameters are set as follows. The maximum degree $d_{\max}=50$, the average degree $d_{\mathrm{ave}}=20$, the maximum community size $c_{\max}=100$, the minimum community size $c_{\min}=20$, the exponents of the power-law distribution of node degrees $\tau_1=2$ and community sizes $\tau_2=1$, respectively.

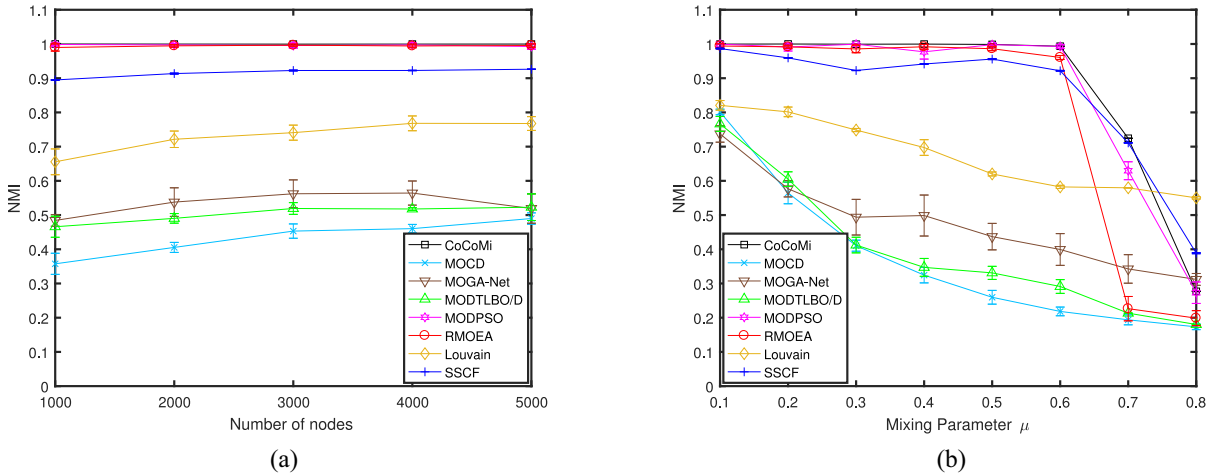


Fig. 8. NMI of the eight algorithms by averaging over 20 runs on different LFR networks. Error bars show the standard deviations estimated from 20 runs. (a) NMI on LFR networks with different numbers of nodes. (b) NMI on LFR networks with different values of μ.

It is worth noting that both fraction of overlapping nodes and number of communities each overlapping node belongs to are fixed as zero, since the compared algorithms were developed to detect communities that do not contain any overlapping node.

We first verify the computational efficiency of the proposed RMOEA on the first two groups of synthetic benchmark networks. Fig. 6(a) and (b) presents the runtime (s) of the eight algorithms by averaging over 20 runs on the first and second groups of LFR networks. From the figures, it can be found that the RMOEA is slightly worse than the two non-EA-based algorithms Louvain and SSCF in efficiency for networks with 1000 to 5000 nodes. Compared with EA-based algorithms, RMOEA performs much better on the LFR networks with 1000 to 5000 nodes in terms of computational efficiency and the superiority will be enhanced as the number of nodes in the network increases. For LFR networks with 5000 nodes, the RMOEA takes less than 40% runtime of all compared EA-based algorithms despite the fact that the runtime of RMOEA will increase as the value of μ increases. The main reason for the runtime increase of RMOEA is attributed to the fact that the size of networks cannot be reduced in the early stage of evolution since the local communities are hard to be identified when the community structure becomes ambiguous as the value of μ increases. To illustrate this fact, Fig. 7 presents the number of nodes in the reduced network after the prereduction (*PreRe*), and the first ($T = 1$) and second ($T = 2$) reductions in evolution by RMOEA averaging over 20 runs on the second group of LFR networks. It can clearly be seen that the size of networks is significantly reduced by the prereduction in case $\mu \leq 0.3$, whereas it is hard to be reduced by the prereduction in case $\mu > 0.3$. However, the size of the networks with $\mu > 0.3$ can be greatly reduced by the first reduction in evolution since many local communities will be found by elite individuals after a few generations. Therefore, the proposed RMOEA is better suited for large-scale LFR benchmark networks than existing EA-based community detection algorithms.

Second, we compare the quality of communities detected by the RMOEA with that of the seven state-of-the-art community detection algorithms. Fig. 8(a) and (b) presents the NMI of the eight algorithms by averaging over 20 runs on the first two groups of LFR benchmark networks. From the figures, we can find that CoCoMi, MODPSO, and the proposed RMOEA achieve the best NMI values on all considered LFR networks whose size ranges from 1000 to 5000 with $\mu = 0.25$. It is also shown that the RMOEA holds a competitive performance on LFR networks when the value of μ varies from 0.1 to 0.6. These empirical results demonstrate the competitiveness of the proposed RMOEA on LFR benchmark networks whose community structure is not very ambiguous. The RMOEA does not achieve a better NMI value than MODPSO and CoCoMi on LFR networks with $\mu > 0.6$. To show the reason, Fig. 9 presents the number of local communities that are correctly and incorrectly reduced in the prereduction (*PreRe*), and the first ($T = 1$) and second ($T = 2$) reductions in evolution by the proposed RMOEA averaging over 20 runs on LFR networks with 5000 nodes and different values of μ. It can be seen that the incorrectly reduced local communities in the first reduction

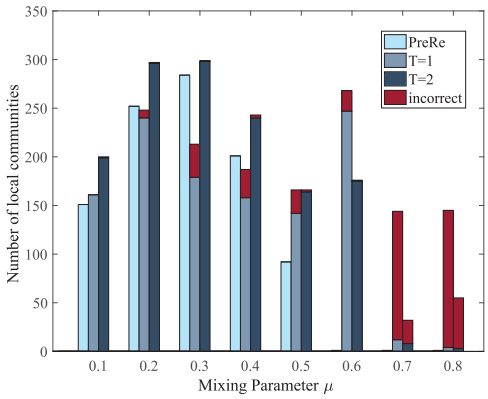


Fig. 9. Number of local communities that are correctly and incorrectly reduced the prereduction (*PreRe*), and the first ($T = 1$) and second ($T = 2$) reductions in evolution by the proposed RMOEA averaging over 20 runs on LFR benchmark networks with 5000 nodes and different values of mixing parameter μ. Note that the correctly reduced local communities mean those whose all nodes belong to a community in the ground truth, while the incorrectly reduced local communities mean those whose at least one node does not belong to the same community in the ground truth.

TABLE III
NMI VALUES AND RUNTIME(S) OF THE EIGHT ALGORITHMS AVERAGING OVER 20 RUNS ON LARGER LFR BENCHMARK NETWORKS

Network	Measure	CoCoMi	MOCD	MOGA-net	MODTLBO/D	MODPSO	SSCF	Louvain	RMOEA
LFR10,000	NMI_avg	/	0.532±0.019$^-$	0.524±0.004$^-$	/	0.974±0.001$^\approx$	**0.996**±0.002$^+$	0.857±0.005$^-$	0.986±0.003
	Runtime	/	6,879	28,533	/	32,561	1,666	63	1,811
LFR20,000	NMI_avg	/	0.588±0.013$^-$	/	/	/	**0.995**±0.001$^\approx$	0.871±0.003$^-$	0.994±0.004
	Runtime	/	20,638	/	/	/	10,430	224	4,326
LFR30,000	NMI_avg	/	0.605±0.018$^-$	/	/	/	/	0.882±0.003$^-$	**0.994**±0.004
	Runtime	/	39,252	/	/	/	/	476	9,490
LFR40,000	NMI_avg	/	/	/	/	/	/	0.889±0.002$^-$	**0.995**±0.004
	Runtime	/	/	/	/	/	/	868	19,331
LFR50,000	NMI_avg	/	/	/	/	/	/	0.894±0.002$^-$	**0.985**±0.004
	Runtime	/	/	/	/	/	/	1,372	39,356
+/-/≈		0/5/0	0/5/0	0/5/0	0/5/0	0/4/1	1/3/1	0/5/0	—

Note that '/' means that NMI values and runtime are not provided here since these results cannot be obtained within 12 hours for one run. The symbols '+', '-' and '≈' indicate that the performance is significantly better, significantly worse and statistically similar to that of RMOEA, respectively.

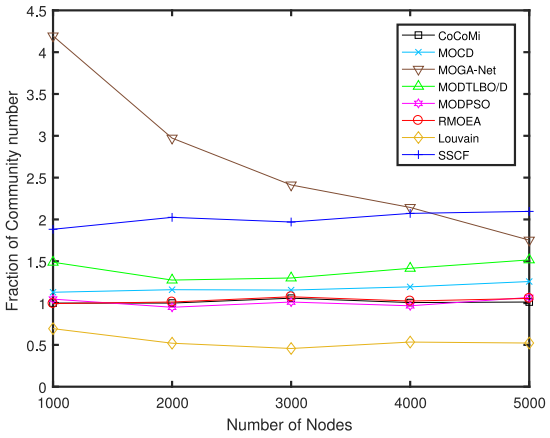


Fig. 10. Fraction of the detected to the known numbers of communities for the eight algorithms on LFR benchmark networks with different numbers of nodes, averaging over 20 networks with the same settings.

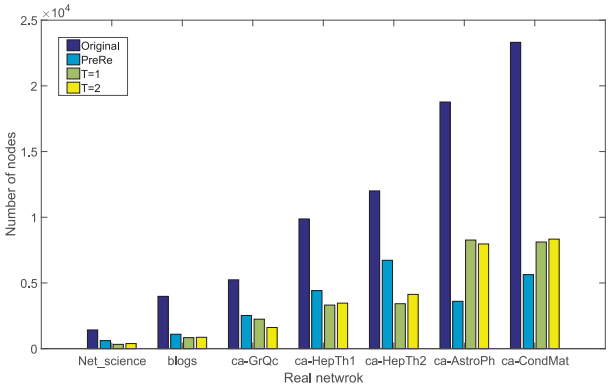


Fig. 11. Number of nodes in the reduced network after the prereduction (*PreRe*), and the first ($T = 1$) and second ($T = 2$) reductions in evolution by the proposed RMOEA averaging over 20 runs on the real-world networks.

in evolution are all corrected in the second reduction when $\mu \leq 0.6$, however, many incorrectly reduced local communities still remain in the second reduction when $\mu > 0.6$ despite the fact that some of them can be corrected. These incorrectly reduced local communities are the main reason why the RMOEA cannot achieve a good performance on LFR networks with $\mu > 0.6$. The competitive performance of the proposed RMOEA can also be observed from Fig. 10, which shows the fraction of the detected to the known numbers of communities for the eight algorithms on LFR benchmark networks with different numbers of nodes.

Third, we test the scalability of the proposed RMOEA in larger networks by using the third group of LFR networks, and the experimental results are listed in Table III. The Wilcoxon rank sum test at a significance level of 0.05 is also adopted to evaluate the statistical difference of the performance of the compared algorithms, in which the symbols "+," "−," and "≈" mean that the result is significantly better, significantly worse, and statistically similar to that obtained by RMOEA, respectively. From this table, we can find that the proposed RMOEA holds a better scalability than the five EA-based algorithms in large-scale networks. Compared with non-EA-based algorithms, the computational efficiency of the proposed RMOEA is worse than Louvain, but is better than SSCF on networks with 20 000 to 50 000 nodes. In terms of the detection performance, the proposed RMOEA statistically shows competitiveness in comparison with the EA-based and non-EA-based algorithms on LFR networks ranging from 10 000 to 50 000 nodes.

Based on the above empirical results, we can conclude that the proposed RMOEA is a promising community detection algorithm for large-scale LFR benchmark networks, in terms of both computational efficiency and detection performance.

C. Experiments on Real-World Networks

In this section, we test the computational efficiency and detection performance of the proposed RMOEA on real-world networks.

Table IV lists the runtime (s) and modularity Q of the eight algorithms for community detection averaging over 20 runs on the seven real-world networks shown in Table II. From the table, it can be found that the non-EA-based algorithm Louvain holds the best computational efficiency on the seven real-world networks and the proposed RMOEA is computationally more efficient than the compared EA-based algorithms since the size of these real-world networks can be significantly reduced by the RMOEA which can be seen from Fig. 11.

In terms of detection performance, the RMOEA achieves a better performance on the real-world networks than on the LFR benchmark networks, despite the fact that the real-world networks were often regarded to hold an ambiguous

TABLE IV
Q VALUES AND RUNTIME(S) OF THE EIGHT ALGORITHMS AVERAGING OVER 20 RUNS ON THE SEVEN REAL-WORLD NETWORKS SHOWN IN TABLE II

Network	Measure	CoCoMi	MOCD	MOGA-net	MODTLBO/D	MODPSO	SSCF	Louvain	RMOEA
Net-science	Q_avg	0.905±0.007$^-$	0.902±0.011$^-$	0.847±0.008$^-$	0.907±0.005$^-$	0.947±0.003$^\approx$	0.830±0.012$^-$	0.894±0.008$^-$	**0.952**±0.004
	Runtime	1,318	379	546	786	995	16	2	321
blogs	Q_avg	0.758±0.006$^-$	0.742±0.008$^-$	0.694±0.012$^-$	0.718±0.004$^-$	0.788±0.005$^-$	0.711±0.005$^-$	0.729±0.007$^-$	**0.806**±0.003
	Runtime	25,657	1,028	2,156	7,965	3,215	3,215	4	628
ca-GrQc	Q_avg	/	0.703±0.007$^-$	0.688±0.008$^-$	0.698±0.002$^-$	0.793±0.015$^\approx$	0.734±0.004$^-$	0.750±0.003$^-$	**0.796**±0.007
	Runtime	/	2,726	3,492	9,682	5,784	369	6	1,272
ca-HepTh1	Q_avg	/	0.460±0.001$^\approx$	**0.486**±0.009$^+$	/	0.462±0.016$^\approx$	0.478±0.001$^-$	0.352±0.001$^-$	0.453±0.005
	Runtime	/	3,835	9,853	/	18,235	836	10	2,465
ca-HepTh2	Q_avg	/	0.405± 0.012$^-$	0.563±0$^-$	/	0.566±0.014$^\approx$	/	**0.583**±0.005$^+$	0.576±0.004
	Runtime	/	9,323	27,864	/	43,168	/	86	3,130
ca-AstroPh	Q_avg	/	0.276± 0.009$^-$	/	/	/	/	0.530±0.004$^-$	**0.535**±0.005
	Runtime	/	21,237	/	/	/	/	305	8,129
ca-CondMat	Q_avg	/	/	/	/	/	/	0.578±0.002$^-$	**0.619**±0.001
	Runtime	/	/	/	/	/	/	116	12,917
+/-/≈		0/7/0	0/6/1	1/6/0	0/7/0	0/3/4	0/7/0	1/6/0	—

Note that '/' means that Q values and runtime are not provided here since these results cannot be obtained within 12 hours for one run. The symbols '+', '-' and '≈' indicate that the performance is significantly better, significantly worse and statistically similar to that of RMOEA, respectively.

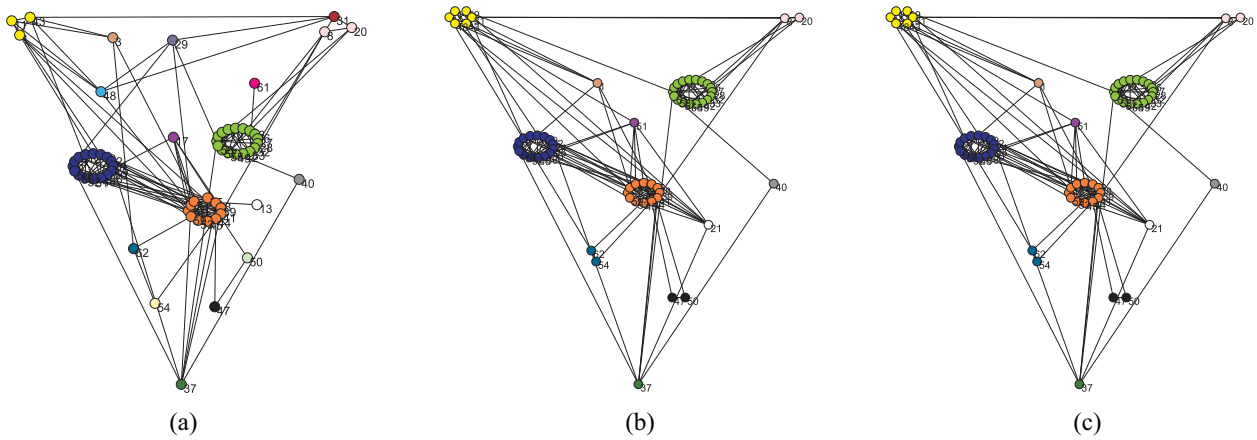


Fig. 12. Reduction results of dolphin network in the prereduction, and the first and second reductions in evolution by the proposed RMOEA. (a) Result after the prereduction, (b) result after the first reduction in evolution, and (c) result after the second reduction in evolution.

community structure. On the seven real-world networks considered here, the proposed RMOEA achieves the best Q value on five networks. These empirical results show that the proposed RMOEA is well suited to detect the community structure in real-world networks. The reason for the RMOEA also achieving a competitive performance on real-world networks with an ambiguous community structure is attributed to the fact that the real-world networks often hold many features which have not been demonstrated in the random benchmark LFR networks. For each of the LFR networks, the degrees of its nodes obey the power-law distributions, which makes LFR networks hard to generate intensive local communities in case that the network has an ambiguous community structure. However, it often occurs that some nodes are closely connected in real-world networks even in case that the community structure of these networks is not distinct. Hence, the real-world networks can still be well reduced by the proposed RMOEA, despite having an ambiguous community structure.

To illustrate this fact, let us take a closer look at the network dolphin, which was first suggested by Lusseau *et al.* [45] for describing the frequent associations between 62 dolphins in a community living of Doubtful Sound, New Zealand. Fig. 12 presents the reduction results of dolphin network in the prereduction, and the first and second reductions in evolution by the proposed RMOEA. As can be seen from the figure, three local communities have been reduced to three nodes by the proposed RMOEA. It is also shown that some nodes of a local community are incorrectly identified in the prereduction, however, they will be corrected in the first and second reduction in evolution by the proposed RMOEA, e.g., the 51-th and 17-th nodes of the dolphin network in Fig. 12.

Therefore, the proposed RMOEA is more suited to detect communities on large-scale real-world networks than the state-of-the-art community detection algorithms, in terms of runtime and detection performance. To further analyze the proposed RMOEA, the effect of prereduction strategy and sensitivity to the number of reductions in RMOEA can be found in the supplementary material IV.

V. CONCLUSION

In this paper, we have proposed a network RMOEA for community detection in large-scale networks. In RMOEA, a

network reduction method has been suggested to recursively reduce the size of the networks in evolution for addressing the curse of dimensionality in large-scale complex networks. The network reduction method consists of two kinds of reductions. The first is conducted before the evolution by using the local topology structure of the network, and the second is done in evolution by using the elite individuals in population. A local community repairing strategy has also been developed in RMOEA to correct the misidentified nodes in the reduced network. Experimental results on synthetic and real-world networks have demonstrated the competitiveness of the proposed RMOEA in community detection for large-scale network, in terms of both computational efficiency and detection performance.

There still remains some work related with RMOEA that deserves to be further investigated. The proposed RMOEA has shown the effectiveness of network reduction in large-scale network for detecting the communities which do not have common nodes. It is interesting to extend the network reduction to overlapping community detection in large-scale networks, since the communities in most real-world networks are overlapped. It is also desirable to consider the network reduction for community detection in other kinds of large-scale complex networks, such as signed networks [33] and dynamical networks [37]. It is worth noting that the calculation of evaluation functions (e.g., modularity Q) for community detection in complex networks is computationally very expensive, which is one of the main reasons for the inefficiency of population-based optimization algorithms, such as EAs. It is interesting to reduce the computational cost by considering the efficient strategies for dealing with expensive problems in EAs, such as surrogate-assisted methods [46]. The effectiveness of RMOEA on networks with millions of nodes also needs to be verified in the future.

REFERENCES

[1] Q. Cai, L. Ma, M. Gong, and D. Tian, "A survey on network community detection based on evolutionary computation," *Int. J. Bio Inspired Comput.*, vol. 8, no. 2, pp. 85–97, 2016.

[2] 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. 1343–1352, 2014.

[3] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York, NY, USA: Cambridge Univ. Press, 1994.

[4] L. A. Adamic *et al.*, "Power-law distribution of the World Wide Web," *Science*, vol. 287, no. 5461, p. 2115, 2000.

[5] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, Jul. 2000.

[6] J. Chen and Y. Saad, "Dense subgraph extraction with application to community detection," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 7, pp. 1216–1230, Jul. 2012.

[7] A. Mahmood and M. Small, "Subspace based network community detection using sparse linear coding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 801–812, Mar. 2016.

[8] C. Pizzuti, "A multi-objective genetic algorithm for community detection in networks," in *Proc. 21st Int. Conf. Tools Artif. Intell.*, 2009, pp. 379–386.

[9] B. Amiri, L. Hossain, and J. W. Crawfordd, "An efficient multiobjective evolutionary algorithm for community detection in social networks," in *Proc. IEEE Congr. Evol. Comput.*, 2011, pp. 2193–2199.

[10] C. Pizzuti, "A multiobjective genetic algorithm to find communities in complex networks," *IEEE Trans. Evol. Comput.*, vol. 16, no. 3, pp. 418–430, Jun. 2012.

[11] C. Shi, Z. Yan, Y. Cai, and B. Wu, "Multi-objective community detection in complex networks," *Appl. Soft Comput.*, vol. 12, no. 2, pp. 850–859, 2012.

[12] M. Gong, L. Ma, Q. Zhang, and L. Jiao, "Community detection in networks by using multiobjective evolutionary algorithm with decomposition," *Physica A Stat. Mech. Appl.*, vol. 391, no. 15, pp. 4050–4060, 2012.

[13] 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.

[14] 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. 1–27, 2015.

[15] D. Chen *et al.*, "Multi-objective optimization of community detection using discrete teaching–learning-based optimization with decomposition," *Inf. Sci.*, vol. 369, no. 10, pp. 402–418, 2016.

[16] D. Corne, N. R. Jerram, J. Knowles, and M. J. Oates, "PESA-II: Region-based selection in evolutionary multi-objective optimization," in *Proc. Genet. Evol. Comput. Conf.*, 2001, pp. 283–290.

[17] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys.*, vol. 11, no. 3, pp. 19–44, 2008.

[18] C. Blundell and Y. W. Teh, "Bayesian hierarchical community discover," in *Proc. Neural Inf. Process. Syst.*, 2013, pp. 1–9.

[19] X. Ma *et al.*, "A multiobjective evolutionary algorithm based on decision variable analyses for multi-objective optimization problems with large scale variables," *IEEE Trans. Evol. Comput.*, vol. 20, no. 2, pp. 275–298, Apr. 2016.

[20] X. Zhang, Y. Tian, R. Cheng, and Y. Jin, "A decision variable clustering-based evolutionary algorithm for large-scale many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 22, no. 1, pp. 97–112, Feb. 2018.

[21] 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.

[22] 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.

[23] C. Pizzuti, "GA-Net: A genetic algorithm for community detection in social networks," in *Proc. Int. Conf. Parallel Problem Solving Nat.*, 2008, pp. 1081–1090.

[24] C. Shi, Z. Yan, Y. Wang, Y. Cai, and B. Wu, "A genetic algorithm for detecting communities in large-scale complex networks," *Adv. Complex Syst.*, vol. 13, no. 1, pp. 3–17, 2010.

[25] H. Chang, Z. Feng, and Z. Ren, "Community detection using dual-representation chemical reaction optimization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4328–4341, Dec. 2017.

[26] 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.

[27] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[28] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 1, pp. 36–41, 2007.

[29] B. Amiri, L. Hossain, and J. W. Crawford, "A hybrid evolutionary algorithm based on HSA and CLS for multi-objective community detection in complex networks," in *Proc. IEEE/ACM Int. Conf. Adva. Soc. Netw. Anal. Min.*, Istanbul, Turkey, 2012, pp. 243–247.

[30] M. Gong, X. Chen, L. Ma, Q. Zhang, and L. Jiao, "Identification of multi-resolution network structures with multi-objective immune algorithm," *Appl. Soft Comput.*, vol. 13, no. 4, pp. 1705–1717, 2013.

[31] B. Amiri, L. Hossain, J. W. Crawford, and R. T. Wigand, "Community detection in complex networks: Multi–objective enhanced firefly algorithm," *Knowl. Based Syst.*, vol. 46, pp. 1–11, Jul. 2013.

[32] F. Zou, D. Chen, S. Li, R. Lu, and M. Lin, "Community detection in complex networks: Multi-objective discrete backtracking search optimization algorithm with decomposition," *Appl. Soft Comput.*, vol. 53, pp. 285–295, Apr. 2017.

[33] C. Liu, J. Liu, and Z. Jiang, "A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2274–2287, Dec. 2014.

[34] A. Amelio and C. Pizzuti, "Community mining in signed networks: A multiobjective approach," in *Proc. Int. Conf. Adv. Soc. Netw. Anal. Min.*, 2013, pp. 95–99.
[35] Q. Cai, M. Gong, B. Shen, L. Ma, and L. Jiao, "Discrete particle swarm optimization for identifying community structures in signed social networks," *Neural Netw.*, vol. 58, no. 10, pp. 4–13, 2014.
[36] K. Kim, R. I. Mckay, and B.-R. Moon, "Multiobjective evolutionary algorithms for dynamic social network clustering," in *Proc. Int. Conf. Genet. Evol. Comput.*, 2010, pp. 1179–1186.
[37] F. Folino and C. Pizzuti, "An evolutionary multiobjective approach for community discovery in dynamic networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1838–1852, Aug. 2014.
[38] Y. Park and M. Song, "A genetic algorithm for clustering problems," in *Proc. 3rd Annu. Conf. Genet. Algorithms*, 1989, pp. 2–9.
[39] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, 2004, Art. no. 066133.
[40] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, 2008, Art. no. 10008.
[41] A. Lancichinetti, S. Fortunato, and F. Radicchio, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, 2008, Art. no. 046110.
[42] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, pp. 2011–2024, 2009.
[43] J. Leskovec and A. Krevl. (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. [Online]. Available: http://snap.stanford.edu/data
[44] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *J. Stat. Mech. Theory Exp.*, vol. 9, no. 1, pp. 1–10, 2005.
[45] D. Lusseau *et al.*, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behav. Ecol. Sociobiol.*, vol. 54, no. 4, pp. 396–405, 2003.
[46] Y. Jin, "Surrogate-assisted evolutionary computation: Recent advances and future challenges," *Swarm Evol. Comput.*, vol. 1, no. 2, pp. 61–70, 2011.

Xingyi Zhang received the B.Sc. degree from Fuyang Normal College, Fuyang, China, in 2003 and the M.Sc. and Ph.D. degrees 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 complex network analysis.

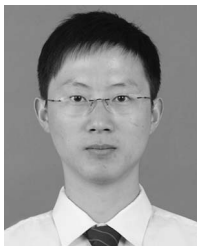
Dr. Zhang was a recipient of the 2017 IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION Outstanding Paper Award.

Kefei Zhou received the B.Sc. degree from Anhui University, Hefei, China, in 2014, where he is currently pursuing the master's degree with the School of Computer Science and Technology.

His current research interest includes multiobjective optimization methods and their application in complex network clustering.

Hebin Pan received the B.Sc. degree from Anhui Architecture University, Hefei, China, in 2012 and the M.Sc. degree from Anhui University, Hefei, in 2017.

His current research interest includes multiobjective optimization methods and their application in complex network clustering.



Lei Zhang received the B.Sc. degree from Anhui Agricultural University, Hefei, China, in 2007 and the Ph.D. degree from the University of Science and Technology of China, Hefei, in 2014.

He is currently a Lecturer with the School of Computer Science and Technology, Anhui University, Hefei. His current research interests include multiobjective optimization and applications, data mining, social network analysis, and pattern recommendation.

Dr. Zhang was a recipient of the ACM CIKM12 Best Student Paper Award.

Xiangxiang Zeng received the B.S. degree in automation from Hunan University, Changsha, China, in 2005 and the Ph.D. degree in system engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2011.

He is currently an Associate Professor with the Department of Computer Science, Xiamen University, Xiamen, China. His current research interests include systems biology, computational intelligence, and data mining.

Yaochu Jin (M'98–SM'02–F'16) received the B.Sc., M.Sc., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1988, 1991, and 1996, respectively, and the Dr.-Ing. degree from Ruhr University Bochum, Bochum, Germany, in 2001.

He is a Professor of computational intelligence with the Department of Computer Science, University of Surrey, Guildford, U.K., where he heads the Nature Inspired Computing and Engineering Group. He is also a Finland Distinguished Professor funded by the Finnish Agency for Innovation (Tekes) and a Changjiang Distinguished Visiting Professor appointed by the Ministry of Education, Beijing, China. He has (co)-authored over 200 peer-reviewed journal and conference papers and been granted eight patents on evolutionary optimization. His current research interests include computational intelligence, computational neuroscience, computational systems biology, and nature-inspired real-world driven problem-solving.

Dr. Jin has delivered 20 invited keynote speeches at international conferences.