Overlapping Community Detection in Directed and Undirected Attributed Networks Using a Multiobjective Evolutionary Algorithm

Xiangyi Teng, Jing Liu, *Senior Member, IEEE*, and Mingming Li

***Abstract*—In many real-world networks, the structural connections of networks and the attributes about each node are always available. We typically call such graphs attributed networks, in which attributes always play the same important role in community detection as the topological structure. It is shown that the very existence of overlapping communities is one of the most important characteristics of various complex networks, while the majority of the existing community detection methods was designed for detecting separated communities in attributed networks. Therefore, it is quite challenging to detect meaningful overlapping structures with the combination of node attributes and topological structures. Therefore, in this article, we propose a multiobjective evolutionary algorithm based on the similarity attribute for overlapping community detection in attributed networks (MOEA-SA$_{OV}$). In MOEA-SA$_{OV}$, a modified extended modularity EQ_{OV}, dealing with both directed and undirected networks, is well designed as the first objective. Another objective employed is the attribute similarity S_A. Then, a novel encoding and decoding strategy is designed to realize the goal of representing overlapping communities efficiently. MOEA-SA$_{OV}$ runs under the framework of the nondominated sorting genetic algorithm II (NSGA-II) and can automatically determine the number of communities. In the experiments, the performance of MOEA-SA$_{OV}$ is validated on both synthetic and real-world networks, and the experimental results demonstrate that our method can effectively find Pareto fronts about overlapping community structures with practical significance in both directed and undirected attributed networks.**

***Index Terms*—Attributed networks, community detection, evolutionary algorithms, multiobjective optimization, overlapping community.**

I. Introduction

NETWORKS arise in many aspects of the real world and provide a powerful tool to represent and analyze various kinds of modern systems [1]. Examples include scientific collaboration networks [2], transportation networks of cities [3], and biological protein–protein interaction networks [4]. Community detection, also called network clustering, is one of the most significant domains of network research, whose aim is to partition all nodes in the network into several clusters such that links are dense within groups but sparse between them [4]–[6]. A good partition of networks always sheds light on the organization and function of systems, which is an important guidance for many fields, such as network visualization and user-targeted online advertising.

Recently, with the rapid growth of information available to us, each node in the graph is always accompanied by one or more attributes describing their properties. Such graphs can be considered as attributed networks, in which node attributes sometimes play the same important role as the topological structure information. As a result, it gives rise to the demand of a new task, attributed network community detection. Much of the recent work focusing on attributed graph clustering targets discovering disjoint groups. Usually, these attributed graph clustering methods can be classified into three categories: 1) distance-based methods [7]; 2) model-based methods [8]; and 3) others [9], [10], [42].

However, as Kelley *et al.* showed in [11], the very existence of overlap is one of the most important characteristics of complex networks, and the phenomena that some nodes in attributed networks belong to several groups often occurs. For instance, in the popular social sites like Facebook, each user could be classified according to different attributes on their profiles, such as education background, profession, and interests. Thus, it poses us a great challenge how to combine and leverage both content information and topological structure to find more relevant and meaningful overlapping communities as well as those influential overlapping nodes.

As we all know, structural connections and attributes of nodes are two totally different types of information, and to some extent, they are even contradictory. Multiobjective evolutionary algorithms (MOEAs) [12], [13], [48], [49] are powerful tools to deal with optimization problems with conflicting objectives. Moreover, in the past few years, MOEAs have been widely applied to community detection problems [9], [10]. However, few clustering methods are intended for the analysis of directed networks where the edge directions contain much potentially useful information. As a matter of fact, directed networks have been around us everywhere, such as the World Wide Web and many social networks. Therefore, there is a great challenge for the researchers to design an algorithm which can deal with overlapping community detection

Manuscript received December 20, 2018; revised April 24, 2019 and July 10, 2019; accepted July 18, 2019. This work was supported in part by the General Program of National Natural Science Foundation of China under Grant 61773300, and in part by the Key Program of Fundamental Research Project of Natural Science of Shaanxi Province, China, under Grant 2017JZ017. This article was recommended by Associate Editor J. Liu. *(Corresponding author: Jing Liu.)*

The authors are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: tengxiangyi@stu.xidian.edu.cn; neouma@mail.xidian.edu.cn; at.mingli@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/TCYB.2019.2931983

problems on both directed and undirected networks. To address the above problems, we propose a multiobjective evolutionary algorithm leveraging attribute information and structural connections for overlapping community detection in both directed and undirected attributed networks (MOEA-SA$_{OV}$). As far as we know, MOEAs have not yet been applied to handle overlapping clustering problems in attributed networks. In MOEA-SA$_{OV}$, a modified extended modularity EQ_{OV}, dealing with both directed and undirected networks, is proposed as the first objective. Another objective employed is attribute similarity S_A. Then, a novel encoding and decoding strategy is designed to realize the goal of representing overlapping communities effectively, which is always the tough part in addressing overlapping community detection problems. MOEA-SA$_{OV}$ runs under the framework of the nondominated sorting genetic algorithm II (NSGA-II) and automatically determines the number of total communities. Four popular measurements, $F1$-score, modularity density D, entropy E, and generalized NMI are employed to evaluate the performance of MOEA-SA$_{OV}$. We conduct experiments on both synthetic networks as well as directed and undirected real-world attributed networks. The experimental results demonstrate that the proposed methods efficiently discover overlapping community structures and determine the number of communities automatically without any prior knowledge. Besides, the overlapping nodes with practical meanings can be also found at the same time. In general, the main contributions of this article are summarized as follows.

1) The multiobjective evolutionary algorithm is first employed on overlapping community detection problems leveraging attribute information and structural connections synchronously. Especially, this algorithm is intended for the analysis of both directed and undirected networks. To this end, we design a new objective called modified extended modularity EQ_{OV} to deal with directed and undirected networks.
2) We propose a novel two-part encoding and decoding strategy to achieve the goal of representing overlapping communities of single individuals without the need to assign the number of communities that each node belongs to in advance. We also adopt the framework of NSGA-II to find a set of Pareto-optimal solutions in a single run for decision-makers to choose according to their different preferences.
3) We conduct extensive experiments on both synthetic networks as well as directed and undirected real-world attributed networks. We also analyze the distribution of community size, overlap size, and node membership obtained by different algorithms at the mesoscopic level. The experimental results in terms of four popular evaluation metrics demonstrate that the proposed methods efficiently discover overlapping community structures and determine the number of communities automatically without any prior knowledge.

The remainder of this article is organized as follows. In Section II, we first introduce the preliminaries to MOEAs and attributed networks as well as the related work on overlapping community detections. Then, two objectives of MOEA-SA$_{OV}$ are described in Section III. In Section IV, we provide a detailed description of our proposed algorithm MOEA-SA$_{OV}$. Section V demonstrates the experimental results obtained by MOEA-SA$_{OV}$ on four synthetic networks and four real attributed networks, together with the comparison with some state-of-the-art algorithms. Finally, in Section VI, we draw a conclusion of the entire article.

II. Preliminaries and Related Work

A. Community Detections in Attributed Networks

For an attributed network G, mathematically, we can regard it as a 3-tuple (V, E, A), where $V = \{v_1, v_2, \ldots, v_n\}$ denotes the set of n nodes, and E denotes the set of edges with $E = \{(v_i, v_j)|v_i \in V, v_i \in V, \text{ and } i \neq j\}$ for undirected networks and $E = \{\overrightarrow{(v_i, v_j)}|v_i \in V, v_j \in V, \text{ and } i \neq j\}$ for directed networks. $A = \{a_1, a_2, \ldots, a_t\}$ is the set of t attributes of each node. The expression $\overrightarrow{(v_i, v_j)}$ means the edge's direction points from node v_i to node v_j. Community detection is a significant domain in network research, whose aim is to partition all nodes in the network into several communities with links being dense within each community but sparse between them. Suppose $\boldsymbol{C} = \{C_1, C_2, \ldots, C_k\}$ is a set of k communities obtained from G, which meets the following requirements:

$$C_i \subset V, \ i = 1, 2, \ldots, k \tag{1}$$

$$C_i \neq \emptyset, \ i = 1, 2, \ldots, k \tag{2}$$

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

$$\bigcup_{i=1}^{k} C_i = V. \tag{4}$$

In order to make communities obtained more practical and meaningful, here, each community in $\boldsymbol{C}$ is considered to be a proper subset of V. Based on the number of communities a vertex belongs to, we can categorize community detection problems as two kinds. If all communities in $\boldsymbol{C}$ satisfy

$$\forall i \neq j \text{ and } i, j \in \{1, 2, \ldots, k\}, \ C_i \cap C_j = \emptyset \tag{5}$$

that is, each node only belongs to one community. Thus, we call it separated community detection problems. Otherwise, if there exists the following situation:

$$\exists i \neq j \text{ and } i, j \in \{1, 2, \ldots, k\}, \ C_i \cap C_j \neq \emptyset \tag{6}$$

then there must be at least one node belonging to more than one community and we call it overlapping community detection problems. In this article, our concentration is basically on how to discover overlapping community structures in attributed networks.

B. Multiobjective Optimization

Mathematically, the problem of community detection can be formally transformed into an optimization problem. Although there are lots of single-objective methods achieved good performance in different kinds of networks [15], [16], the intuitive notion of a meaningful and reasonable community in attributed networks should consider two competing parts: 1) structural connections and 2) node attribute information. Thus, it is natural to model the problem as a multiobjective optimization problem.

Let $\Omega = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_u\}$ be the set of all candidate solutions for community detection, then we can define a multiobjective problem as

$$F(\boldsymbol{C}^*) = \min_{\boldsymbol{C}\in\Omega} F_i(\boldsymbol{C}), \;\; i = 1, 2, \ldots, p \tag{7}$$

where $\boldsymbol{F} = \{F_1, F_2, \ldots, F_p\}$ is a set of p objective functions, and some functions may conflict with each other. Therefore, we may not reach the minimum values on all functions simultaneously. Multiobjective optimization methods are applied to obtain a solution C^* so that every conflicting part in objective functions could have a pleasant tradeoff value. Here, we use the concept of nondominated solutions or Pareto-optimal solutions [12]–[14] to define the optimal solution. To be specific, suppose we have two feasible solutions, denoted by X_1 and X_2. X_1 dominates X_2 when and only when X_1 is not worse than X_2 on all objectives and better on one or more objectives. For multiobjective optimization problems, the set of solutions usually contains several nondominated solutions. When all the nondominated solutions are plotted in the objective space, they are termed as Pareto fronts (PFs). The solutions on a PF often serve as candidate choices for decision-makers to choose according to different situations.

C. Related Work

The past few decades have witnessed the great progress in developing methods to uncover community structure of different kinds of networks. Fortunato [6] made a thorough exposition of different methods and categorized them into corresponding types. However, many community detection methods above focus on either link analysis or content information and, thus, the resultant clustering cannot reflect both topological structure and node properties in the input graph inevitably.

For attributed networks, several approaches have been proposed recently to combine structural connections and vertex information for a good and meaningful partition. These attempts can be divided into three types: 1) distance-based methods; 2) model-based methods; and 3) others. As for distance-based methods [7], they usually employ a distance measure which takes both the structure and attribute information into consideration. Therefore, these methods can adjust the weights iteratively according to such measures until they strike a good balance between structural connections and attribute information. As for the model-based approaches [8], [17], [43], these methods simulate the process of generating networks according to different Bayesian probabilistic models in order to make use of structural and attribute information naturally and theoretically. Moreover, there exist some work on attribute similarity for defining the optimization objective [9], [42]. Lately, Li *et al.* [9] introduced a multiobjective evolutionary algorithm based on attributed similarities to address clustering problems in attributed graphs. In the proposed algorithm, attribute similarity S_A was used to measure how the homogenous of the nodes inside a cluster is. Besides, the methods of SAC1 and SAC2 [42] define attribute similarity based on the Euclidean distance. Each algorithm mentioned above achieved a relatively good result on tested benchmarks and real-world networks, however, most of the methods cannot be employed to detect overlapping communities.

With more studies and researches conducted on network community structure, phenomena often occur that some nodes in attributed networks have multiple memberships. To address this issue, there is a surge of interest in developing community detection algorithms for discovering communities which are not necessarily disjoint in the past few years. These algorithms can be categorized into three classes, namely, the methods based on clique percolation [18], link partitioning, and genetic algorithm.

The clique percolation methods are based on the typical k-clique algorithm, whose key idea is that communities are likely to be made up of small cliques sharing many of their nodes with other cliques. Then, a k-clique community is regarded as the union of all adjacent k-cliques that share $k-1$ nodes. However, this kind of methods is more like pattern matching rather than detecting communities because they tend to find small and localized structure of a network. As a result, some overlapping nodes and communities may be missed.

The link partitioning approaches employ the idea of discovering links rather than nodes that participate in several communities. Ahn *et al.* [19] proposed a method partitioning links according to edge similarity. Then a link dendrogram is created by single-linkage hierarchical clustering. Finally, we can cut this dendrogram at some threshold to get the final link overlapping communities. Although the concept of link partitioning methods seems natural, they cannot guarantee better performance over node-based approaches because they are based on an ambiguous definition of communities.

Pizzuti [20] proposed a single-objective genetic algorithm called GA-NET+ to detect overlapping communities. The proposed method operates on the line graph $L(G)$ and optimizes a fitness function called community score to get a better partition of networks. With the rapid development of MOEAs, some researchers begin to employ MOEAs for overlapping community detection. Liu *et al.* [25] proposed an MOEA called MEA_CDPs which is based on a permutation representation scheme for detecting separated and overlapping communities. Wen *et al.* [33] proposed an MOEA based on maximal-clique graph for overlapping community detection.

In many real-world systems, graphs are directed with the source node transmitting some properties to the target one. Therefore, revealing the underlying community structure of directed networks becomes a significant topic recently. Basically, the approaches can be divided into three categories. The first and simplest way is to discard edge direction and treat them as undirected. The approaches in the second category transform directed graph into an undirected one, while information and semantics about the direction of the edges are meaningfully incorporated [46]. The third category is extending objective functions and methodologies to be suitable for directed networks [47].

Most of the aforementioned works deal with either overlapping community detection or attributed network clustering. Our focus in this article is on the intersection of both two domains. Moreover, our proposed algorithm is designed for dealing with both directed and undirected networks. MOEAs are powerful tools for dealing with optimization problems with

competing objectives [12], [13], [38]. Thus, taking the advantages of MOEAs, we propose a multiobjective evolutionary algorithm leveraging attribute information and structural connections for overlapping graph clustering in both directed and undirected attributed networks (MOEA-SA$_{OV}$).

III. Two Objectives of MOEA-SA$_{OV}$

For community detection problems, one of the most well-known functions to evaluate clustering results is modularity Q [21]. Formally, modularity Q is expressed as

$$Q = \frac{1}{2m}\sum_{q=1}^{k}\sum_{i\in C_q, j\in C_q}\left(A_{ij} - \frac{d_i d_j}{2m}\right) \tag{8}$$

where k represents the number of communities, A denotes the adjacency matrix of the original network, m represents the total number of edges, and d_i and d_j denote the degrees of nodes v_i and v_j, respectively. Later, a new version of modularity, EQ, used for overlapping community detection was proposed in [22]

$$EQ = \frac{1}{2m}\sum_{q=1}^{k}\sum_{i\in C_q, j\in C_q}\frac{1}{O_i O_j}\left(A_{ij} - \frac{d_i d_j}{2m}\right) \tag{9}$$

where O_i and O_j denote how many communities that nodes v_i and v_j belong to. What should be noted is that EQ fits well for undirected networks but may perform unpleasantly for directed networks due to the ambiguous and inappropriate definition of d_i. For a directed network, the total degree of each node equals to the sum of its in-degree and out-degree. It is inappropriate to only employ the total degree of a node in overlapping community detection. Thus, here, we come up with a modified EQ_{OV} suitable for both directed and undirected graphs as the first objective to be maximized

$$EQ_{OV} = \frac{1}{2m}\sum_{q=1}^{k}\sum_{i\in C_q, j\in C_q}\frac{1}{O_i O_j}\left(A_{ij} - \frac{d'_i d'_j}{2m}\right) \tag{10}$$

$$di' = \begin{cases} \left|\left\{v_j \middle| (v_i, v_j) \in E \text{ and } i \neq j\right\}\right| & \text{if } G \text{ is undirected} \\ \left|\left\{v_j \middle| \overrightarrow{(v_j, v_i)} \in E \text{ and } i \neq j\right\}\right| & \text{if } G \text{ is directed} \end{cases} \tag{11}$$

where we take d'_i as the total degree of node v_i for undirected networks and in-degree of node v_i for directed networks.

As for the second objective, we use the attribute similarity S_A proposed in [9] to evaluate how homogenous the nodes inside clusters of G are. S_A naturally makes full use of the attribute information of vertices, which is a complementation of the network topological structure. S_A also needs to be maximized, and can be expressed as

$$SA = \frac{\sum_{q=1}^{k}\sum_{vi,vj\in C_q, i<j} 2s(i,j)}{\sum_{q=1}^{k} r_q(r_q - 1)} \tag{12}$$

in which r_q denotes how many vertices inside community C_q and $s(i, j)$ calculates the similarity between nodes v_i and v_j according to their attribute values. Usually, an attribute value can be categorized as discrete or continuous. We only take discrete attribute into consideration in this article. Considering a discrete attribute $b \in A$, the definition of $s(i, j)$ is expressed as

$$s(i,j) = \begin{cases} 1 & b_i = b_j \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where b_i and b_j are the attribute values of nodes v_i and v_j in terms of attribute b.

In this article, the discussion is mainly about single-attribute networks. However, the proposed methods can be easily applied to deal with multiple attribute networks by redefining $s(i, j)$ as the cosine similarity

$$s(i,j) = \frac{\overrightarrow{i}\cdot\overrightarrow{j}}{\left|\overrightarrow{i}\right| \times \left|\overrightarrow{j}\right|} \tag{14}$$

where $\overrightarrow{i}$ and $\overrightarrow{j}$ are the sets of attributes of nodes v_i and v_j.

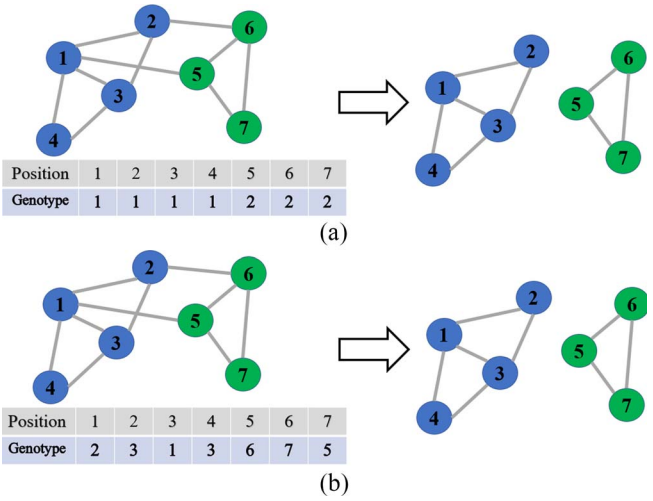

Fig. 1. (a) Label-based and (b) locus-based representations of a given network containing seven vertices and ten links.

IV. MOEA-SA$_{OV}$

A. Hybrid Direct and Indirect Representation

The representation of EAs plays a significant role in the convenience of employing evolutionary operators to EAs and the efficiency of an algorithm. According to the existing literature of EAs, representations can be divided into two broad types: 1) direct representation and 2) indirect representation. As for the direct representation, the label-based [23] and locus-based representations [24] have been proved to be quite efficient. Let a genotype be denoted as an integer vector of size $n, X = [x_1, x_2, \ldots, x_n]$, and each position in the vector corresponds to a node. For the label-based representation, suppose there is totally k communities, then every gene x_i randomly takes a number from integer set $\{1, 2, \ldots, k\}$ as the label representing the cluster it participates in. While in the locus-based representation, each gene x_i can have an allele value j in the set $\{1, 2, \ldots, n\}$. A value j assigned to x_i indicates there is a connection between nodes v_i and v_j and they are exactly in the same cluster. Fig. 1 shows the detailed procedure about both representations for a given graph.

As for the indirect representation, we cannot just use it to complete the task of encoding the division of a network.

Algorithm 1 Decoding Method

Input:
$X = \{x_1, x_2, \ldots, x_n\}$: genotype in which each position represents a node;
k: total number of communities after the locus-based representation is used;
$C' = \{C'_1, C'_2, \ldots, C'_k\}$: a set of separated communities obtained after the first initialization part;
Output:
$C = \{C_1, C_2, \ldots, C_k\}$: a set of overlapping communities;

```
begin
    C = C';
    for i = 1 to n do
    begin
        for j = 1 to k do
        begin
        if (x_i ∉ C'_j) then
            if (F(C_j) < F(C_j ⋃ x_i)) then
                C_j ← C_j ⋃ x_i;
                F(C_j) = F(C_j ⋃ x_i);
        end
    end
end
```

Usually, we require a decoder to help the entire representation process. In our previous work [25], we employ a heuristic search, which actually is a process of optimizing the community fitness function proposed in [26] to represent the overlapping communities. Suppose C_i be a community, and the community fitness function of C_i can be defined as

$$F(C_i) = \frac{k_{\text{in}}^{C_i}}{\left(k_{\text{in}}^{C_i} + k_{\text{out}}^{C_i}\right)^{\alpha}} \tag{15}$$

where $k_{\text{in}}^{C_i}$ and $k_{\text{out}}^{C_i}$ denote the internal and external degrees of all nodes in C_i, respectively, and α here is a resolution parameter. For the sake of simplicity, we set $\alpha = 1$ in this article.

Borrowing ideas from the above works, we design a novel encoding and decoding strategy to realize the goal of representing overlapping communities effectively. The strategy includes two parts. In the initialization part, to make full use of the adjacent structural information for a better initialization, the direct locus-based adjacency representation is employed to the initial population. This encoding step is more efficient and accelerates the process of evolution compared with the permutation-based representation in [25]. Then, we transform them into label-based representation where it is much easier to employ evolutionary operators. It should be noted that at this time each node already belongs to one community. In all the following evolutionary process, we use the label-based representation.

Since the label-based representation cannot be applied to overlapping community detection problems, we design the second part which is an indirect decoding process, where we adopt the community fitness function described in [26] to achieve the goal of representing overlapping communities of single individual. Fig. 2 gives an illustrative example of decoding process. Every node could join other communities as long as the value of fitness functions increasing. The two parts above together constitute our hybrid direct and indirect representation. Algorithm 1 shows the details of our decoding method.

TABLE I
PROCEDURE OF TWO-WAY CROSSOVER OPERATION WHEN SELECTING v_4

v	X_a	X_b	X_c	X_d	X_a	X_b	v
1	③	6	③	3	3	6	1
2	1	1	1	1	1	1	2
3	7	4	4	④	7	④	3
④	③	4	③	④	3	④	④
5	1	6	6	1	1	6	5
6	③	4	③	④	3	④	6
7	4	3	3	4	4	3	7

B. Evolutionary Operators

1) Crossover Operator: For the label-based representation, some classic operators like one-point operator might create invalid solutions where nodes having no edge between them could be partitioned into the same cluster. Therefore, these traditional operators cannot be applied to our method. A novel two-way crossover operator has shown great efficiency in [27], so we adopt this operator through the evolutionary process. The detailed crossing over procedure is described in Table I. First, we use the binary tournament selection to select X_a and X_b as parent chromosomes. Then, one vertex v_i (node 4 in Table I) is randomly chosen whose community label is denoted as k_a (community 3 in X_a). Make sure that all nodes in community k_a of X_a are also assigned to the same community in X_b. At the same time, suppose k_b (community 4 in X_b) represents the community label of vertex v_i in X_b. Ensure that all nodes in this community of X_b are also assigned to the same community k_b in X_a. After conducting such a process, two children chromosomes (X_c and X_d) are generated.

2) Multi-Individual-Based Mutation Operator: The mutation operators are adopted to change gene values so that the search process could go toward the regions not yet inspected. Usually, it can hardly get a better solution with the action of randomly modifying a value. Here, we apply a heuristic method proposed in our previous work [9], which is called multi-individual-based mutation. The basic idea of this operator is to update a target individual taking advantages of the useful information from three existing individuals. We consider three situations based on the relationship among their community labels in a certain position. Then, a neighborhood correction strategy is applied to fix improper genes if exist. The more detailed steps can be referred to [9].

C. Implementation of MOEA-SA$_{OV}$

NSGA-II [28] has long been regarded as a significant EA for multiobjective optimization problems. This algorithm generates nominated sets with different levels and provides evenly distributed solutions (for specific description of NSGA-II refer to [28]). Therefore, MOEA-SA$_{OV}$ is implemented under the framework of NSGA-II.

Algorithm 2 shows the details of MOEA-SA$_{OV}$. First, we randomly initialize a population with M chromosomes using the locus-based representation. Then, each chromosome is transformed into the character string representation and we have a set of separated communities for now. Next, we apply the decoding method described in Algorithm 1 to get a set of overlapping communities. The part of representing

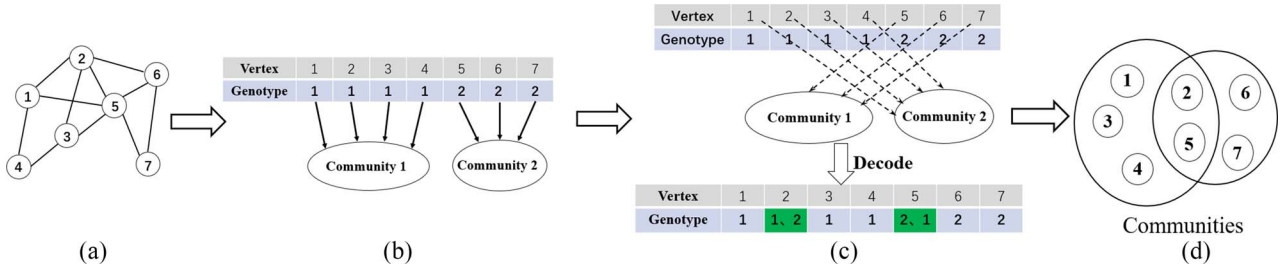

Fig. 2. Illustrative examples of the decoding process (schema). (a) Original graph. (b) One possible genotype after the initialization (each vertex belongs to only one community now). (c) One possible genotype after the decoding process (here, community fitness function is the criterion for each vertex to join other communities). (d) Obtained communities according to the genotype after the decoding process.

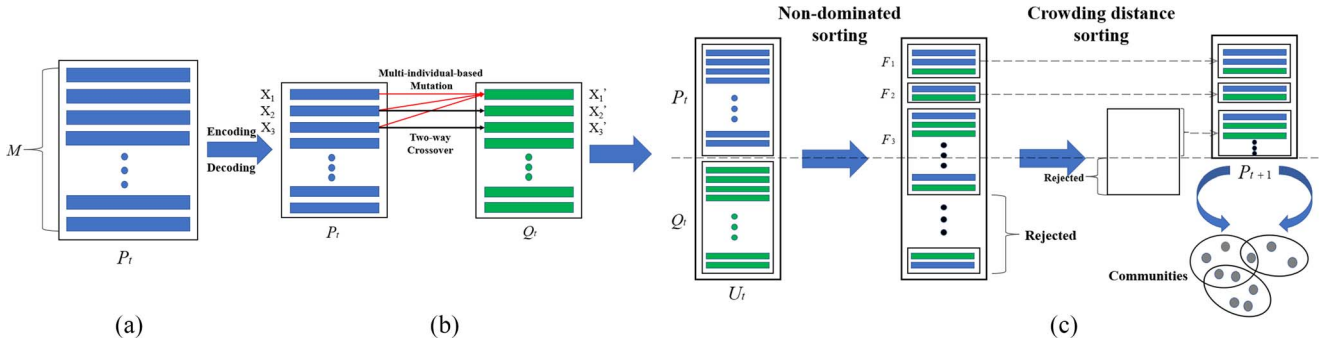

Fig. 3. Flowchart of MOEA-SA$_{OV}$. (a) Initial parent population containing M individuals from generation t. (b) Parent population P_t after conducting the encoding and decoding process, then the binary-tournament-selection operator, multi-individual-based mutation operator and two-way crossover operator are applied on the parent population to generate offspring generation Q_t with the size of M. (c) U_t is the combination of parent and offspring populations. Nondominated sorting and crowding distance sorting are employed to get the best M individuals from U_t to form a new parent population P_{t+1} based on the values of fitness functions. Iteratively upgrade populations until the termination condition is reached.

overlapping communities is finished. Now, we calculate two objective functions EQ_{OV} and S_A to measure the quality of each chromosome. In the next stage, the crossover operator and mutation operator are applied to create offspring populations. The new offspring population is decoded into a set of overlapping communities according to Algorithm 1 and evaluated by two objective functions. Finally, the elite strategy is conducted to create the parent population for the next generation. Iteratively run the above steps before the termination condition of MOEA-SA$_{OV}$ is reached. For an intuitional understanding of the procedures, we also give the flowchart of MOEA-SA$_{OV}$ in Fig. 3.

D. Complexity Analysis

Let n and m be the number of nodes and edges in graph G. *Popsize* denotes how many individuals in a single population and $G_{\max}$ represents the maximum number of generations. The computational complexity of our algorithm is determined by three stages. For the initialization stage, the computational complexity is $O(\text{Popsize} \times n)$. Next, for each individual, both of the crossover and mutation operators take $O(n)$ operations, and the complexity of decoding stage is $O(n \times k)$, where k is the number of communities after applying the locus-based representation. For the function evaluation stage, the computational complexity is $O(m + n)$. In most of the cases where there are more edges than nodes in a network, the complexity of this stage could be $O(m)$. So, the computational complexity of MOEA-SA$_{OV}$ can be simplified to $O(G_{\max} \times \text{Popsize} \times (m + n))$.

V. Experiments

In this section, we conduct varieties of experiments on both synthetic and real-world networks to verify the performance of MOEA-SA$_{OV}$. All experiments are conducted on a 4-core personal computer with a 3.3-GHz CPU and 8-GB memory. MOEA-SA$_{OV}$ is implemented in C++. The size of population is 100 and the maximum number of generations is 50. All experiments are conducted under the same parameter setting.

A. Datasets Description

We use both synthetic benchmark networks and real-world networks in our experiments. Synthetic networks are generated by the LFR benchmark [35], and real-world networks include two directed and two undirected networks. The detailed descriptions about these networks are given as follows.

1) LFR Benchmark Networks [35]: All synthetic networks are generated by the LFR benchmark with ten parameters to be set: 1) number of nodes n; 2) average degree $\overline{d}$; 3) maximum degree *maxd*; 4) mixing parameter μ, which denotes the fraction of edges of a node linking to other communities; 5) power-law degree distribution with exponent t_1; 6) power-law community size distribution with exponent t_2; 7) minimum for the community size *minc*; 8) maximum for the community size *maxc*; 9) number of overlapping nodes O_n; and 10) number of memberships of the overlapping nodes O_{mem}. According to the spirit of the comparison performed in [36] and [37], we design four typical synthetic networks with different characteristics to validate the efficiency of MOEA-SA$_{OV}$. To make the comparison be fair, the same parameters for four networks

Algorithm 2 MOEA-SA$_{OV}$

Input:
Popsize: number of individuals in a single population;
G$_{max}$*:* maximum number of generations;
p_c*:* crossover probability;
$G = (V, E, A)$: attribute network;
Output:
A set containing all the non-dominated solutions;
Series of Symbols:
P_t*:* parent population containing *popsize* individuals from the tth generation;
Q_t*:* offspring population containing *popsize* individuals from the tth generation;
U_t*:* combination of parents and offspring populations containing 2× *Popsize* individuals from the tth generation;
F*:* a non-dominated set that consists of a variety of rank levels;

$t \leftarrow 1$;
initialize P_t with the locus-based representation, then transform P_t into the label-based representation;
For each individual in P_t, use the decoding method in Algorithm 1 to obtain a set of overlapping communities, and calculate the two objective functions;
while ($t <$ G$_{max}$) **do**
  $Q_t \leftarrow \emptyset$;
  while ($|Q_t| <$*Popsize*) **do**
    $[X_1, X_2, X_3] \leftarrow$ binary-tournament-selection (P_t);
    $X'_1 \leftarrow$multi-individual-based-mutation (X_1, X_2, X_3);
    $[X'_2, X'_3] \leftarrow$crossover-operator (X_2, X_3, p_c);
    $Q_t \cup Q_t \cup [X'_1, X'_2, X'_3]$;
  end while
For each individual in Q_t, use the decoding method in Algorithm 1 to obtain a set of overlapping communities, and calculate the two objective functions;
  $U_t \leftarrow P_t \cup Q_t$;
  $F \leftarrow$ fast-nondominated-sort (U_t);
  $P_{t+1} \leftarrow \emptyset$;
  $i \leftarrow 1$;
  while ($|P_{t+1}| + |F_i| <$*Popsize*) **do**
    $P_{t+1} \leftarrow P_{t+1} \cup F_i$;
    $i \leftarrow i + 1$;
  end while
  Calculate the crowding distance of each individual in F_i;
  Sort F_i in descending order according to the crowding distance;
  $P_{t+1} \leftarrow P_{t+1} \cup F_i$ [1:(*Popsize* -$|P_{t+1}|$)];
  $t \leftarrow t + 1$;
end while

are set as follows: $\overline{d} = 5$, max $d = 25$, $t_1 = 2$, $t_2 = 1$, min $c = 20$, max $c = 80$, and $O_n = 30\%$ of the total number of nodes. The details of the different parameters for these four synthetic networks are shown in Table II.

2) American College Football Network [4]: It is a network collected form college football games of USA taking place in the autumn of 2000. There are total 12 different attribute values for each node indicating the universities they are on behalf of.

3) Political Books Network [29]: This network collects political books of USA published around 2004 national presidential election. All books are purchased from the website Amazon.com. If there is a link between two books, it means customers frequently buy these books all together. For every book, there is a single attribute showing their political tendencies for three choices: 1) conservative; 2) liberal; or 3) neutral.

4) WebKB Network [30]: This network is composed of 877 scientific publications and 1608 connections, which includes webpage networks of four American universities: 1) Cornell; 2) Washington; 3) Texas; and 4) Wisconsin. Each university contains 195, 230, 187, and 265 nodes, respectively. Each webpage network has an attribute with five values: 1) course; 2) faculty; 3) student; 4) project; and 5) stuff. It is a directed network with the head of the link pointing to the webpages being cited and the tail of the link pointing to the webpages which contain the citation. In the following part, we conduct our experiments on two of the four universities' webpage networks, namely, Cornell and Washington.

TABLE II
PARAMETER SETTING OF FOUR SYNTHETIC NETWORKS

Network	n(size)	μ	O_{mem}
LFR1	1000	0.1	2
LFR2	1000	0.1	3
LFR3	1000	0.2	2
LFR4	5000	0.1	2

B. Evaluation Measurements

For synthetic networks with the ground-truth community structure, we use $F1$-*score* as a measure of accuracy to quantify the ability of MOEA-SA$_{OV}$ to detect overlapping nodes. $F1$-*score* is defined as

$$F1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{16}$$

where *precision* denotes the ratio of correctly detected overlapping nodes to the total detected overlapping nodes and *recall* equals to the ratio of correctly detected overlapping nodes to the ground-truth overlapping nodes. $F1$-*score* is the harmonic mean of *precision* and *recall*.

For real-world networks, two popular metrics, namely, density D and entropy E, are employed to evaluate the performance of MOEA-SA$_{OV}$

$$D = \sum_{q=1}^{k} \frac{m_q}{m} \tag{17}$$

where m_q denotes how many links in community C_q, m represents the total number of edges, and k stands for the number of clusters. The density D, to some extent, is the reflection of the proportion of community intralinks over the total number of edges in the network. The higher the density D is, the more evident the obtained community structures are.

Let the set of values of attribute b be $\{1, 2, \ldots, l\}$, and entropy E is expressed as

$$E = \sum_{q=1}^{k} \frac{r_q}{n} \cdot \text{entropy}(q) \tag{18}$$

$$\text{entropy}(q) = -\sum_{i} p_{iq} \cdot \log\left(p_{iq}\right) \tag{19}$$

where r_q denotes how many nodes in community C_q, n denotes the number of nodes in graph, and p_{iq} is the percentage of nodes in community C_q whose attribute values are i. The obtained clusters with smaller E means that the nodes inside clusters are more homogenous.

To further verify the performance of MOEA-SA$_{OV}$, generalized normalized mutual information (gNMI) [26] is used

TABLE III
COMPARISON RESULTS IN TERMS OF *F1-Score* ON FOUR LFR BENCHMARK NETWORKS WITH DIFFERENT CHARACTERISTICS

Network	Metric	MOEA-SA$_{OV}$	CFinder	SLPA
LFR1	*F1-score*	**0.4634**	0.2609	0.3852
	recall	**0.7600**	0.1700	0.4333
	precision	0.3333	**0.5604**	0.3467
LFR2	*F1-score*	**0.4454**	0.1698	0.4441
	recall	**0.7267**	0.1033	0.567
	precision	0.3211	**0.4769**	0.365
LFR3	*F1-score*	**0.4311**	0.1150	0.4022
	recall	**0.6833**	0.0667	0.4967
	precision	0.3149	**0.4167**	0.3379
LFR4	*F1-score*	**0.4495**	0.2388	0.4203
	recall	**0.7193**	0.1473	0.4693
	precision	0.3221	**0.6296**	0.3805

to estimate the similarity between the network with true community structure and a detected one. gNMI (A, B) is expressed as

$$\mathrm{gNMI}(A,B)=\frac{-2\sum_{i=1}^{k_A}\sum_{j=1}^{k_B}R_{ij}\log\left(R_{ij}\cdot n/R_{i.}R_{.j}\right)}{\sum_{i=1}^{k_A}R_{i.}\log(R_{i.}/n)+\sum_{j=1}^{k_B}R_{.j}\log\left(R_{.j}/n\right)} \tag{20}$$

where n represents the total number of nodes in the network, and R denotes the confusion matrix of which element R_{ij} records the number of nodes in community C_i from division A that are also in community C_j from division B. k_A (k_B) denotes the number of communities in division $A(B)$, $R_{i.}$ ($R_{.j}$) stands for the sum of the elements from row i (column j). A larger value of gNMI means our algorithm could obtain a community structure more similar to the ground truth.

C. Experimental Results on the Synthetic Networks

1) Performance Evaluation in Terms of F1-Score: In this experiment, to test the performance of MOEA-SA$_{OV}$ on four typical LFR benchmark networks, we compare MOEA-SA$_{OV}$ with two popular methods, namely, CFinder [18], which is an implementation of clique percolation methods, and SLPA [31], which is a label propagation algorithm. The time complexity of CFinder is polynomial in many applications [18]. SLPA has $O(\mathrm{Tm})$ time complexity on an arbitrary network, where T is the maximum number of iterations and m is the number of edges in this network. We record the result with the best setting for algorithms with tunable parameter. Specifically, in CFinder, k varies from 3 to 6. Parameter r of SLPA varies from 0.1 to 0.5 in the step of 0.05. Here, *F1-score* is employed as an evaluation metric to quantify the accuracy and efficiency of the algorithms above. Detailed experimental results are shown in Table III.

Table III shows that MOEA-SA$_{OV}$ outperforms the other algorithms in terms of *F1-score* and *recall* on networks with different characteristics. It is important to find that MOEA-SA$_{OV}$ tends to detect more correct overlapping nodes due to its novel encoding and decoding strategy and fitness function design. However, detecting more overlapping nodes also leads to a relatively low *precision.* Although CFinder has higher *precision* on all the four networks, it can only detect quite a small number of overlapping nodes, and the overall performance of CFinder is relatively poor.

2) Investigation of Mesoscopic Properties of the Networks: At the mesoscopic level, Palla *et al.* [18] introduced four measures to quantify the overlapping community structures of networks. In this experiment, we make the comparison in terms of three of them, namely, the distribution of community size, the overlap size, and the node membership obtained by different algorithms. Here, for the sake of clarity, we take synthetic network LFR3 as an example. We compare MOEA-SA$_{OV}$ with two popular algorithms CFinder [18] and SLPA [31]. Figs. 4–6 present the distribution of community size, overlap size, and node membership obtained by different algorithms, respectively.

As can be observed from Figs. 4–6, all distributions are close to the power-law distribution. However, compared with the two other algorithms, MOEA-SA$_{OV}$ is the best fit of power-law distribution considering all these three mesoscopic properties. From Figs. 5(b) and 6(b), we can find that CFinder tends to detect less nonoverlapping nodes as well as overlapping nodes in the obtained community structure.

3) Performance Evaluation in Terms of gNMI: In this experiment, to further test the performance of MOEA-SA$_{OV}$ on synthetic networks, we use gNMI as an evaluation metric. Here, for the sake of clarity, we also take synthetic network LFR3 as an example. We compare MOEA-SA$_{OV}$ with CFinder and SLPA. The maximum gNMI (denoted as gNMI_max) as well as the average gNMI (denoted as gNMI_avg) are recorded for each algorithm. Fig. 7 shows the detailed comparison results.

As can be seen from Fig. 7, MOEA-SA$_{OV}$ achieves the better performance on the synthetic network LFR3 in terms of both maximum and average gNMI, which highlights the good performance of MOEA-SA$_{OV}$.

D. Pareto Fronts Obtained by MOEA-SA$_{OV}$

Because our approach is a multiobjective evolutionary algorithm, all the results obtained are a series of nondominated solutions. To verify the good performance of MOEA-SA$_{OV}$, we plot the final PFs for each network obtained by our method, which are shown in Fig. 8. As can be seen, MOEA-SA$_{OV}$ clearly obtains enormous nondominated solutions and quite evenly distributed PFs.

From Fig. 8, we can obviously observe that nearly all the nondominated solutions obtained in four real networks have EQ_{OV} greater than 0.3. In Fig. 8(a) and (b), even the minimum $EQov$ values are both greater than 0.4. Usually, a network whose EQ_{OV} more than 0.3 is credited with community structure. In addition, when EQ_{OV} is close to 0.9, S_A in Polbook is greater than 0.75, which means more than three quarters of the nodes inside clusters share the same attribute value. Based on the above analysis, each solution in PFs is acceptable with cohesive community structure and true homogeneity. Moreover, all PFs in Fig. 8 indicate that EQ_{OV} and S_A are indeed two competing objectives that one increases its value at the cost of loss of the other.

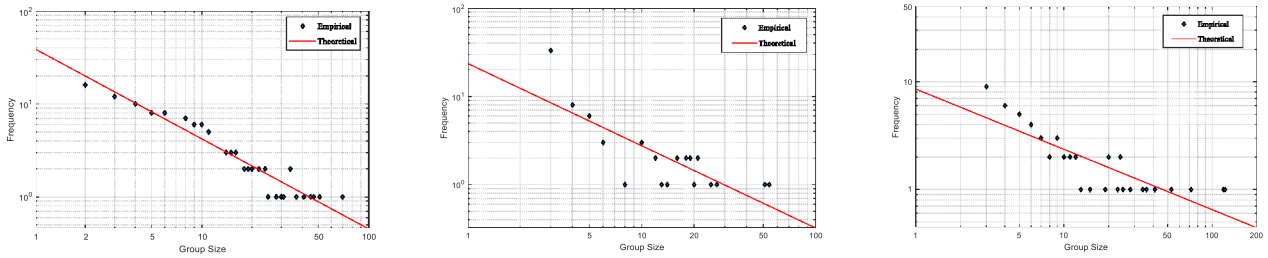

Fig. 4. Log–log empirical group size distribution (dots) of network LFR3 and power-law estimation (line). (a) MOEA-SA$_{OV}$. (b) CFinder. (c) SLPA.

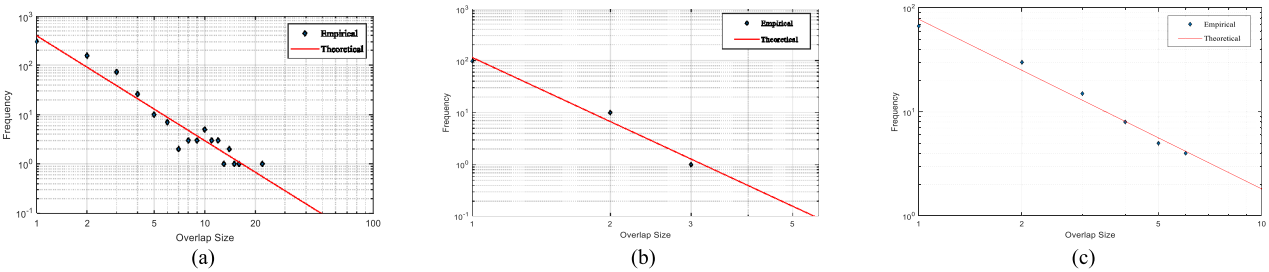

Fig. 5. Log–log empirical overlap size distribution (dots) of network LFR3 and power-law estimate (line). (a) MOEA-SA$_{OV}$. (b) CFinder. (c) SLPA.

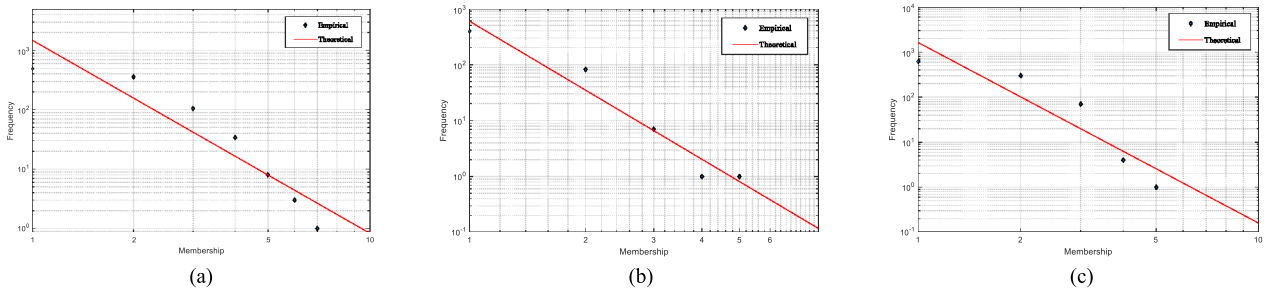

Fig. 6. Log–log empirical membership distribution (dots) of network LFR3 and power-law estimate (line) of network LFR3. (a) MOEA-SA$_{OV}$. (b) CFinder. (c) SLPA.

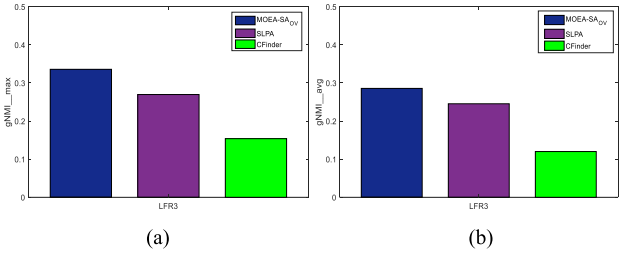

Fig. 7. Detailed comparison results in terms of (a) maximum gNMI and (b) average gNMI on the synthetic network LFR3.

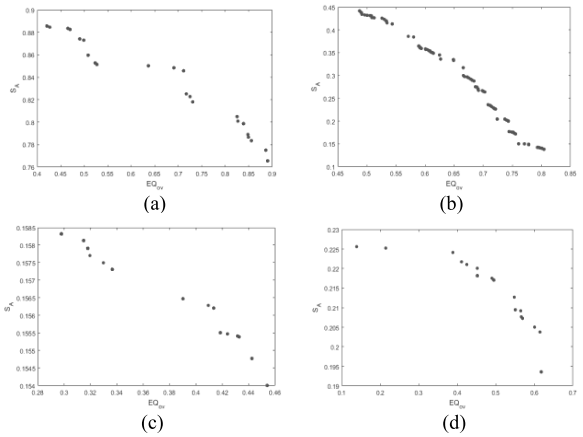

Fig. 8. PFs obtained by MOEA-SA$_{OV}$ for real-world attributed networks. (a) Political Books. (b) American College Football. (c) Cornell. (d) Washington.

E. Experimental Results in Terms of D and E

In this section, we first study the relationship between D and E as the number of communities varying. Then, we compare our method with CFinder and SLPA in terms of D and E. Unlike many existing algorithms that require setting the number of communities in advance, MOEA-SA$_{OV}$ provides a diverse range of communities for different requirements in practice. Decision makers are able to flexibly choose a solution according to specific occasions. The number of communities in each network obtained by MOEA-SA$_{OV}$ is shown in Table IV. Further, in each network, the relationship between D/E (average value of the final generation) and the number of overlapping communities (O_m) is illustrated in Fig. 9.

From Fig. 9, it is interesting to find that Density D and Entropy E have the same variation trends when the number

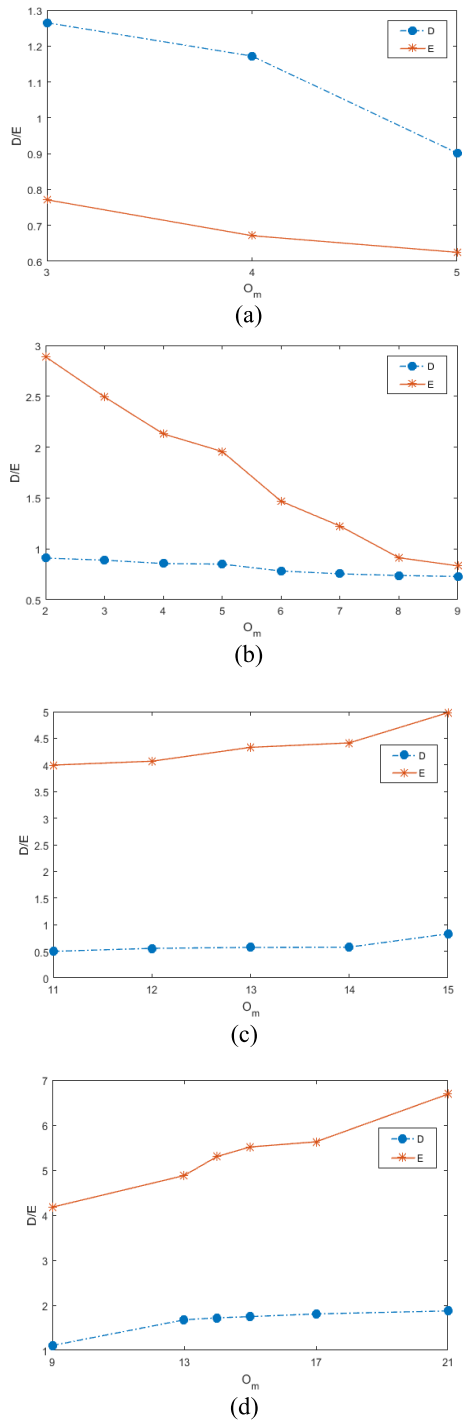

Fig. 9. Variation tendencies of D and E as O_m increasing for four real networks. (a) Political Books. (b) American College Football. (c) Cornell. (d) Washington.

of detected communities increases. Specifically, in Fig. 9(a) and (b), both D and E go down as the number of clusters increases. In Fig. 9(c) and (d), both D and E decrease when the number of communities increases. Since the larger value of D reflects more obvious community structure and the smaller value of E means vertices inside clusters tend to be more homogenous, the same variation trend means the relationship between D and E is conflicting. A solution with larger D cannot have smaller E at the same time. To some great extent, it also proves that topological structure (related to the first objective EQ_{OV}) and node attribute (related to the second objective S_A) are complementary information for community detection. Considering only one of them would possibly fail to get practical and significant structure of networks.

TABLE IV
NUMBER OF COMMUNITIES OBTAINED BY MOEA-SA$_{OV}$

Network	MOEA-SA$_{OV}$
Political books	3-5
American College Football	2-9
Cornell	11-15
Washington	9-21

TABLE V
COMPARISON RESULTS IN TERMS OF D AND E ON REAL-WORLD NETWORKS

Network	Metric	MOEA-SA$_{OV}$	CFinder	*SLPA*
Political books	D	**1.1202**	0.9932	0.9728
	E	**0.6230**	0.7733	0.9320
American football	D	0.9322	**0.9560**	0.8858
	E	2.2230	3.2493	**1.8485**
Cornell	D	**0.8350**	0.4033	0.8167
	E	4.6250	**0.8903**	1.6812
Washington	D	**1.6768**	0.4861	1.0127
	E	4.8791	**0.6455**	1.8807

Next, to further verify the performance of MOEA-SA$_{OV}$, we also compare MOEA-SA$_{OV}$ with CFinder and SLPA in terms of D and E. We run the experiments ten times and records the result with the best setting for algorithms with tunable parameter. Specifically, in CFinder, k varies from 3 to 6. In SLPA, parameter r varies from 0.01 to 0.3 in the step of 0.05. Since MOEA-SA$_{OV}$ attains many nondominated solutions in a single run, we select the solutions from PFs with the closest values of D or E compared with that of the two other algorithms. Table V shows the detailed experimental results.

As we can see from Table V, MOEA-SA$_{OV}$ achieves relatively good results in terms of D and E. Especially, MOEA-SA$_{OV}$ obtains the highest value of D in almost all networks. For the political books network, MOEA-SA$_{OV}$ performs better in terms of both D and E with 1.1202 and 0.6230, respectively. However, MOEA-SA$_{OV}$ tends to get greater value of E in some networks, which leads to the solution with nodes inside clusters are not so homogeneous. The reason why this phenomenon occurs is probably that MOEA-SA$_{OV}$ tends to find more communities sometimes.

F. Experimental Results in Terms of gNMI for Real-World Networks

To further validate the advantages of MOEA-SA$_{OV}$, we additionally use gNMI as an evaluation metric. gNMI originates from the information science and is widely used in network analysis. However, gNMI can only be applied

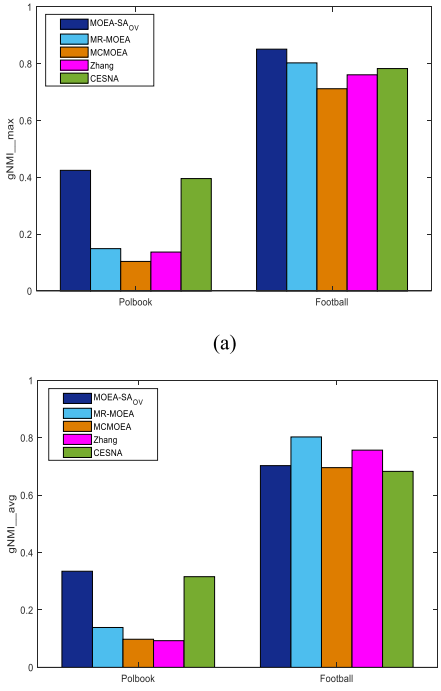

Fig. 10. Detailed comparison in terms of (a) maximum gNMI and (b) average gNMI on the Political Books and American College Football network.

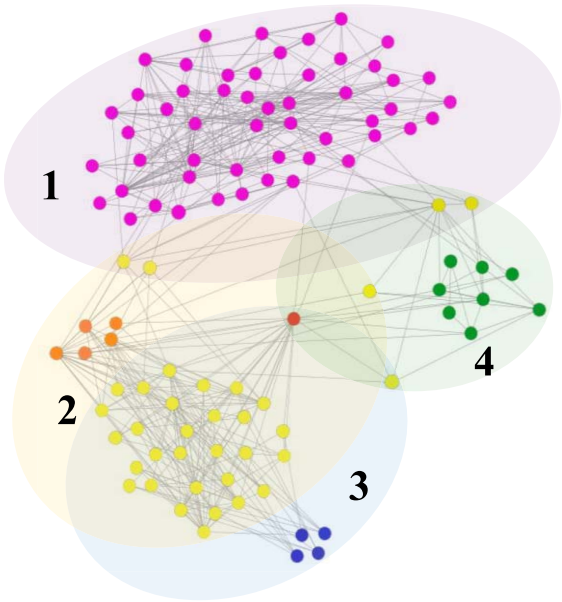

Fig. 11. Illustrative example of the community structure about the Political Books network obtained by MOEA-SA$_{OV}$. There are four communities in total represented by different colors, namely, pink, green, orange, and blue. All of the yellow nodes are overlapping nodes involved in two communities, and the red one in the center is partitioned into three communities.

to networks with ground-truth community structure, therefore, political books network [29] and American football network [4] are used in this part.

For the comparison algorithms, two recently proposed MOEA-based overlapping community detection algorithms, namely, MR-MOEA [32] and MCMOEA [33], together with two famous algorithms that do not belong to MOEAs, namely, Zhang's algorithm [34] and CESNA [43], are also selected. The computational complexities of MR-MOEA, MCMOEA, and CESNA are the same as that of MOEA-SA$_{OV}$. Zhang's algorithm has exponential time complexity, which is higher than our method. For a fair comparison, two MOEA baseline approaches use the same size and maximum number of generations as MOEA-SA$_{OV}$. The tunable parameters of each algorithm are set as the suggestion of the corresponding paper. As for Zhang's algorithm [34], we adopt the code and parameters provided by their authors. We record the maximum gNMI (denoted as gNMI_max) as well as the average gNMI (denoted as gNMI_avg) of each algorithm. Fig. 10 gives the detailed comparison.

As shown in Fig. 10, we can clearly see that MOEA-SA$_{OV}$ can obtain much better results in terms of gNMI (NMI_max) on the Political Books and American College Football networks. With regard to average gNMI, our method gets the best result on Political Books network while performs worse than MR-MOEA and Zhang's algorithm on the American College Football network. Since MOEA-SA$_{OV}$ gets larger number of communities on this network, it is acceptable that the value of average gNMI is not so great. Based on all the analysis above, we could draw a conclusion that MOEA-SA$_{OV}$ shows the competitive performance compared with other start-of-the-art methods on attributed network community detection.

G. Visualization of Overlapping Nodes and Community Structure

In this part, to better illustrate the good performance of MOEA-SA$_{OV}$, we visualize the community structures obtained by MOEA-SA$_{OV}$ on two networks. One is an undirected real-world network, namely, the Political Books network, and the other is a directed real-world network, namely, the Corporate Law Partnership Advice network [50] which contains 36 partners from a Northeastern U.S. corporate law firm. In this network, each partner is described by various attributes and here we consider their law school which has three values: 1) Harvard or Yale; 2) UConn; and 3) others. The directed edge x pointing to y links two layers whenever x tends to go to y for basic professional advice. The visualizations of the two networks are shown in Figs. 11 and 12, respectively.

As can be seen from Fig. 11, MOEA-SA$_{OV}$ obtains four overlapping communities in the Political Books network. It is worth noticing that most of the detected overlapping nodes participate two communities. The only exception is the node colored in red, which has three memberships since it has much higher degree and almost the same number of edges linked to those three communities. This phenomenon consists with the conclusion in [40] that the diversity of overlapping nodes in social networks is relatively small, typically 2 or 3. Moreover, we can find that more than half of the nodes in communities 2 and 3 are overlapping nodes since they represent the books purchased together with all the books denoted by nodes colored in blue and orange frequently. And because the nodes colored in blue have no links with the nodes colored in orange, they naturally belong to different communities.

Fig. 12 presents a visualized community structure obtained by MOEA-SA$_{OV}$ on the Corporate Law Partnership Advice network. Our algorithm obtains three communities in total

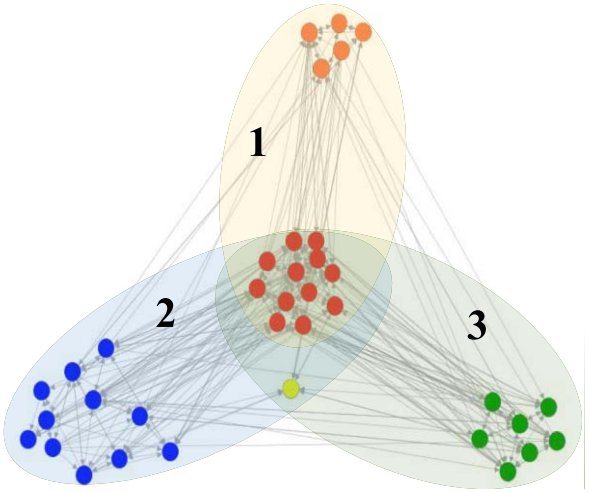

Fig. 12. Illustrative example of the community structure about the Corporate Law Partnership Advice network obtained by MOEA-SA$_{OV}$. There are three communities in total represented by different colors, namely, orange, blue, and green. All of the red nodes are overlapping nodes involved in three communities, and the yellow one is partitioned into two communities.

which are represented by different colors. The yellow node in the figure is the overlapping node which is involved in two communities and the nodes with overlap size more than two are reparented in red. It is interesting to find that most of the detected overlapping nodes participate into three communities since basically they all act as the bridge to connect the three communities and their in-degrees from separated nodes are relatively high. Moreover, most of the separated nodes in three communities are the partners from the same law school, which shows that, to some extent, partners in this firm tend to ask for advice from people in the same law school. Similar to the phenomenon shown in Fig. 11, the overlapping nodes with multiple memberships are usually the ones which have higher degree and almost the same number of edges linked to the communities they belong to.

VI. Conclusion

In this article, a novel multiobjective evolutionary algorithm, called MOEA-SA$_{OV}$, is proposed for overlapping community detection in attributed networks. In our algorithm, the modified extended modularity EQ_{OV}, dealing with both directed and undirected networks, is proposed as the first objective. Another objective employed is attribute similarity S_A. Especially, a novel encoding and decoding strategy is designed to realize the goal of representing overlapping communities effectively. MOEA-SA$_{OV}$ runs under the framework of NSGA-II and strikes great balance between topological structure and vertex properties. Four popular evaluation measurements are used to verify the quality of communities obtained by MOEA-SA$_{OV}$. The performance of MOEA-SA$_{OV}$ is tested on four synthetic networks as well as two directed and two undirected real attributed networks, and the experimental results demonstrate that MOEA-SA$_{OV}$ can efficiently discover overlapping community structures and determine the number of communities automatically without any prior knowledge. Besides, all overlapping nodes with practical meanings can be also found at the same time. In the future, our top priority will be put on implementing the algorithm in the parallel pattern and use the surrogate model [44], [45], [52] as well as the techniques from estimation of the distribution algorithms [51] to reduce the computational time in detecting community structures of networks with hundreds of thousands of nodes. We would also like to further study more fast and effective encoding and decoding schemes for a better representation. Moreover, multiple criteria decision-making strategies, such as the knee point-driven method [39] and group decision-making approach [41], can be also combined with this article.

References

[1] M. E. J. Newman, *Networks: An Introduction*. Oxford, U.K.: Oxford Univ. Press, 2010.
[2] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.
[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
[4] 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.
[5] Z. Jiang, J. Liu, and S. Wang, "Traveling salesman problems with PageRank distance on complex networks reveal community structure," *Physica A Stat. Mech. Appl.*, vol. 463, pp. 293–302, Dec. 2016.
[6] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.
[7] Y. Zhou, H. Cheng, and J. X. Yu, "Clustering large attributed graphs: An efficient incremental approach," in *Proc. Int. Conf. Data Min.*, 2010, pp. 689–698.
[8] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "A model-based approach to attributed graph clustering," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Scottsdale, AZ, USA, 2012, pp. 505–516.
[9] Z. Li, J. Liu, and K. Wu, "A multiobjective evolutionary algorithm based on structural and attribute similarities for community detection in attributed networks," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 1963–1976, Jul. 2018.
[10] 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.
[11] S. Kelley, M. Goldberg, M. Magdon-Ismail, K. Mertsalov, and A. Wallace, "Defining and discovering communities in social networks," in *Handbook of Optimization in Complex Networks* (Springer Optimization and Its Applications), vol. 57, M. Thai and P. Pardalos, Eds. Boston, MA, USA: Springer, 2012, pp. 139–168.
[12] C. M. Fonseca and P. J. Fleming, "Multiobjective optimization and multiple constraint handling with evolutionary algorithms. I. A unified formulation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 28, no. 1, pp. 26–37, Jan. 1998.
[13] N. Srinivas and K. Deb, "Multiobjective optimization using nondominated sorting in genetic algorithms," *Evol. Comput.*, vol. 2, no. 3, pp. 221–248, 1994.
[14] H. Ishibuchi and T. Murata, "A multi-objective genetic local search algorithm and its application to flowshop scheduling," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 3, pp. 392–403, Aug. 1998.
[15] 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.
[16] C. Pizzuti, "GA-NET: A genetic algorithm for community detection in social networks," in *Proc. 10th Int. Conf. Parallel Problem Solving Nat.*, 2008, pp. 1081–1090.
[17] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "GBAGC: A general Bayesian framework for attributed graph clustering," *ACM Trans. Knowl. Disc. Data*, vol. 9, no. 1, pp. 1–43, 2014.
[18] 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.
[19] 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.

[20] C. Pizzuti, "Overlapped community detection in complex network," in *Proc. Genet. Evol. Comput. Conf. (GECCO)*, Montreal, QC, Canada, 2009, pp. 859–866.

[21] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, 2004, Art. no. 026113.

[22] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A Stat. Mech. Appl.*, vol. 388, no. 8, pp. 1706–1712, 2009.

[23] M. Tasgin, A. Herdagdelen, and H. Bingol, "Community detection in complex networks using genetic algorithms," *arXiv:0711.0491 [physics.soc-ph]*, 2007.

[24] Y. J. Park and M. S. Song, "A genetic algorithm for clustering problems," in *Proc. 3rd Annu. Conf. Genet. Program.*, 1998, pp. 568–575.

[25] J. Liu, W. Zhong, H. A. Abbass, and D. G. Green, "Separated and overlapping community detection in complex networks using multiobjective evolutionary algorithms," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Barcelona, Spain, 2010, pp. 1–7.

[26] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure of complex networks," *New J. Phys.*, vol. 11, no. 3, 2009, Art. no. 033015.

[27] M. Gong, B. Fu, L. Jiao, and H. Du, "Memetic algorithm for community detection in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 84, no. 5, 2011, Art. no. 056101.

[28] 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.

[29] V. Krebs. (2004). *Books About U.S. Politics*. [Online]. Available: http://www.orgnet.com

[30] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, pp. 93–106, 2008.

[31] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *Proc. IEEE 11th Int. Conf. Data Min. Workshops (ICDMW)*, 2011, pp. 344–349.

[32] 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.

[33] 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.

[34] T. Zhang and B. Wu, "A method for local community detection by finding core nodes," in *Proc. Int. Conf. Adv. Soc. Netw. Anal. Min.*, 2012, pp. 1171–1176.

[35] A. Lancichinetti, S. Fortunato, and F. Radicchi, "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.

[36] G. K. Orman, V. Labatut, and H. Cherifi, "Comparative evaluation of community detection algorithms: A topological approach," *J. Stat. Mech. Theory Exp.*, no. 8, 2012, Art. no. p08001.

[37] M. Jebabli, H. Cherifi, C. Cherifi, and A. Hamouda, "Community detection algorithm evaluation with ground-truth data," *Physica A Stat. Mech. Appl.*, vol. 492, pp. 651–706, Feb. 2018.

[38] 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, Aug. 2015.

[39] X. Zhang, Y. Tian, and Y. Jin, "A knee point-driven evolutionary algorithm for many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 19, no. 6, pp. 761–776, Dec. 2015.

[40] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surveys*, vol. 45, no. 4, 2013, Art. no. 43.

[41] Y. Wu, Y. Dong, J. Qin, and W. Pedrycz, "Flexible linguistic expressions and consensus reaching with accurate constraints in group decision-making," *IEEE Trans. Cybern.*, to be published.

[42] T. A. Dang and E. Viennet, "Community detection based on structural and attribute similarities," in *Proc. Int. Conf. Digit. Soc.*, Valencia, Spain, 2012, pp. 7–12.

[43] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proc. 13th IEEE Int. Conf. Data Mining*, Dallas, TX, USA, 2013, pp. 1151–1156.

[44] H. Wang, Y. Jin, C. Sun, and J. Doherty, "Offline data-driven evolutionary optimization using selective surrogate ensembles," *IEEE Trans. Evol. Comput.*, vol. 23, no. 2, pp. 203–216, Apr. 2019.

[45] M. O. Akinsolu, B. Liu, V. Grout, P. I. Lazaridis, and M. E. Mognaschi, "A parallel surrogate model assisted evolutionary algorithm for electromagnetic design optimization," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 2, pp. 93–105, Apr. 2019.

[46] V. Satuluri and S. Parthasarathy, "Symmetrizations for clustering directed graphs," in *Proc. 14th Int. Conf. Extend. Database Technol.*, 2011, pp. 343–354.

[47] Y. Kim, S.-W. Son, and H. Jeong, "Finding communities in directed networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 81, Jan. 2010, Art. no. 016103.

[48] W.-N. Chen, Y.-H. Jia, F. Zhao, X.-N. Luo, X.-D. Jia, and J. Zhang, "A cooperative co-evolutionary approach to large-scale multisource water distribution network optimization," *IEEE Trans. Evol. Comput.*, to be published.

[49] Q. Yang *et al.*, "A distributed swarm optimizer with adaptive communication for large-scale optimization," *IEEE Trans. Cybern.*, to be published.

[50] E. Lazega, *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford, U.K.: Oxford Univ. Press, 2001.

[51] Q. Yang, W. Chen, Y. Li, P. Chen, X. Xu, and J. Zhang, "Multimodal estimation of distribution algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 636–650, Mar. 2017.

[52] H. Wang and Y. Jin, "A random forest-assisted evolutionary algorithm for data-driven constrained multiobjective combinatorial optimization of trauma systems," *IEEE Trans. Cybern.*, to be published.

Xiangyi Teng received the B.S. degree in electronic and engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree in circuits and systems with the School of Artificial Intelligence, Xidian University, Xi'an, China.

His current research interests include complex networks and evolutionary algorithms.

Jing Liu (SM'15) received the B.S. degree in computer science and technology and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2000 and 2004, respectively.

In 2005, she joined Xidian University as a Lecturer, and was promoted to a Full Professor in 2009. From 2007 to 2008, she was with the University of Queensland, Brisbane, QLD, Australia, as a Postdoctoral Research Fellow, and from 2009 to 2011, she was with the University of New South Wales at the Australian Defence Force Academy, Canberra, ACT, Australia, as a Research Associate. She is currently a Full Professor with the School of Artificial Intelligence, Xidian University. Her current research interests include evolutionary computation, complex networks, fuzzy cognitive maps, multiagent systems, and data mining.

Prof. Liu is an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. She was the Chair of Emerging Technologies Technical Committee of IEEE Computational Intelligence Society from 2017 to 2018.

Mingming Li received the B.S. degree in intelligence science and technology from Xidian University, Xi'an, China, in 2017, where he is currently pursuing the M.S. degree in circuits and systems with the School of Artificial Intelligence.

His current research interests include machine learning, evolutionary algorithms, and complex networks.