2021 IEEE Congress on Evolutionary Computation (CEC) | 978-1-7281-8393-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CEC45853.2021.9504894

A Reduced Mixed Representation Based Multi-Objective Evolutionary Algorithm for Large-Scale Overlapping Community Detection

Yongkang Luo*
School of Computer Science
and Technology
Anhui University
Hefei, China
luoyongkang@stu.ahu.edu.cn

Kening Zhang*
School of Computer Science
and Technology
Anhui University
Hefei, China
Kirnie@163.com

Haipeng Yang
School of Computer Science
and Technology
Anhui University
Hefei, China
haipengyang@126.com

Feng Liu
School of Computer Science
and Technology
Anhui University
Hefei, China
liufeng0194@outlook.com

Shuai Luo
School of Computer Science
and Technology
Anhui University
Hefei, China
2109760579@qq.com

Lei Zhang
School of Computer Science
and Technology
Anhui University
Hefei, China
zl@ahu.edu.cn (*corresponding author*)

Xiaoyan Sun
School of Information and
Control Engineering
China University of Mining and Technology
Xuzhou, China
xysun78@126.com

***Abstract*—In recent years, the application of multi-objective evolutionary algorithms (MOEAs) to overlapping community detection in complex networks has been a hot research topic. However, the existing MOEAs for detecting overlapping communities show poor scalability to large-scale networks due to the fact that the encoding length of individuals is usually equal to the number of all nodes in the network. To this end, we suggest a reduced mixed representation based multi-objective evolutionary algorithm named RMR-MOEA for large-scale overlapping community detection, where the length of the individual is recursively reduced as the evolution proceeds. Specifically, a mixed representation is adopted for fast encoding and decoding the individual in the population, which consists of two parts: one represents all potential overlapping nodes and the other represents all non-overlapping nodes. Then, in each individual length reduction, two strategies are suggested to respectively shorten the length of each part in the mixed representation, with the aim to greatly reduce the search space. Finally, the experimental results on 10 real-world complex networks demonstrate the effectiveness of the proposed RMR-MOEA in terms of both detection performance and running time, especially on large-scale networks.**

***Index Terms*—large-scale complex network, multi-objective optimization, evolutionary algorithm, overlapping community detection, mixed representation.**

I. Introduction

Community detection is a very vital tool for uncovering the information on all kinds of complex systems in a number of domains, such as the internet network [1], biological network [2], and social network [3]. Specifically, the task of community detection is to divide a network into several groups of nodes (i.e. communities) based on the topology structure of the network, where nodes in the same community have a tight connection while nodes in different communities have a sparse connection [4]. Thus, this task can be formulated as a multi-objective optimization problem by simultaneously maximizing the number of internal links in communities and minimizing the number of external links between different communities [5]. To this end, designing effective multi-objective evolutionary algorithms (MOEAs) for community detection in complex networks has attracted a large number of researchers, due to the fact that MOEAs can return a set of Pareto optimal solutions for multiple selections and overcome some potential disadvantages such as the limited resolution of modularity.

Among the existing MOEA-based community detection algorithms, many researchers devote themselves to designing non-overlapping community detection algorithms, where each one node must belong to one and only one community [6]–[13]. However, some nodes in real-world communities often belong to two or more communities. For example, in a scientist collaboration network, one person might be a member of “machine learning” community and “evolutionary computation” community simultaneously. In a social network, one person might be a member of “football” community as well as a member of “basketball” community. To this end, researchers began to focus on designing overlapping community detection algorithms based on MOEAs, where each node in real-world networks may belong to two or more communities.

For example, in 2010, Liu *et al.* [14] proposed an MOEA based algorithm named MEA-CDPs to detect separated and overlapping communities simultaneously. After that, they extended MEA-CDPs for signed network and proposed another

[*] Authors contributed equally to this work.

978-1-7281-8393-0/21/$31.00 ©2021 IEEE

MOEA based algorithm named MEAs-SN for detecting both separated and overlapping communities [15]. However, the search space of these two algorithms increases exponentially as the number of nodes in the network goes up. In 2015, Li *et al.* [16] developed an improved multi-objective quantum-behaved particle swarm optimization named IMOQPSO on the basis of spectral clustering, which was verified on the small scale real-world and synthetic networks. In 2017, Wen *et al.* [17] proposed a maximal clique based MOEA, termed MC-MOEA, for detecting overlapping communities. However, the large number of communities and the long individual length may result in the performance degradation. In 2017, Zhang *et al.* [18] proposed a mixed representation based MOEA named MR-MOEA for detecting overlapping communities, where a mixed individual representation scheme was designed to fast encode and decode the overlapping division of networks. In 2020, Tian *et al.* [19] proposed a multi-objective evolutionary optimization based fuzzy method (named EMOFM) for overlapping community detection. However, the time complexity of the algorithm is still high, since it has to find overlapping nodes by optimizing fuzzy threshold of every node in the network.

Recently, Ma *et al.* [20] proposed a local-to-global scheme based MOEA (named LG-MOEA) for overlapping community detection on large-scale complex networks. Specifically, LG-MOEA consists of two stages, that is, a local community structure detection stage and a community structure determination stage. In the first stage, an MOEA with the proposed community boundary control strategy was suggested to detect the multiple possible local community structures instead of directly detecting the global community partition on the whole network, thus LG-MOEA can deal with large-scale networks. Then, in the second stage, the global overlapping community partition of the whole network was determined by a single objective EA.

Experimental results on different complex networks have demonstrated the superiority of these MOEA-based overlapping community detection methods over traditional ones. However, these MOEA-based algorithms still show poor scalability to large-scale networks because of the curse of dimensionality, since the individual length of encoding a network is equal (or proportional) to the number of nodes in the network. In other words, the individual representation length of these algorithms remains stable as the evolution proceeds. To this end, different from the above MOEAs, in this paper, we propose a reduced mixed representation based multi-objective evolutionary algorithm named RMR-MOEA for large-scale overlapping community detection, where the length of individual is recursively reduced as the evolution proceeds.

To be specific, the mixed representation proposed in [18] is adopted for rapidly encoding and decoding the network divisions, where the nodes are classified into potential overlapping nodes and non-overlapping nodes. Then, in each length reduction of individual, the historical information is used to fix potential overlapping nodes, while the local communities that existed in elite individuals are used to shorten non-overlapping nodes. These two strategies can be used in RMR-MOEA to greatly reduce the search space. Finally, the effectiveness and efficiency of the proposed RMR-MOEA are verified on 10 real-world networks, and the experimental results demonstrate that RMR-MOEA is superior over several representative baseline algorithms for overlapping community detection, especially on large-scale networks.

II. The Preliminaries

In this section, we first give some preliminaries about overlapping community detection problem, and then present the adopted mixed-representation scheme for decoding overlapping communities.

A. Multi-Objective Overlapping Community Detection

In this paper, we consider only undirected and unweighted complex network. We use a graph denoted as $G = (V, E)$ to represent a network, where $V = \{v_1, v_2, \cdots, v_n\}$ denotes the set of all nodes in G and $E = \{(i, j) | v_i \in V, v_j \in V \ and \ i \neq j\}$ denotes the set of all edges in G. Given a network G, the task of overlapping community detection is to divide the nodes in G into a set of communities or groups, where each node maybe belong to two or more communities. Formally, the set of all detected communities in G is denoted as $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, where C_i satisfies the following conditions:

$$C_i \subset V \ and \ C_i \neq \emptyset, i = 1, 2, ..., k \quad (1)$$

$$C_i \neq C_j, \forall i \neq j \ and \ i, j \in \{1, 2, ..., k\} \quad (2)$$

$$C_i \cap C_j \neq \emptyset, \exists i \neq j \ and \ i, j \in \{1, 2, ..., k\} \quad (3)$$

$$\bigcup_{i=1}^{k} C_i = V \quad (4)$$

Note that each community is a proper subset of V and the joint set of all communities is equal to V.

In the community detection problem, nodes in the same community have a tight connection while nodes in different communities have a sparse connection. To this end, the community detection problem in EAs can be modeled as a multi-objective optimization problem with two conflicting objectives [8], [16], [17]. To be specific, the first objective is to maximize the *intra-link* density, that is, link density between nodes in the same community. The other one is to minimize the *inter-link* density, that is, link density between nodes in different communities. Note that, several criteria are proposed for measuring *intra-link* and *inter-link* densities. In this paper, the kernel k-means (KKM) [3] is adopted for measuring *intra-link* density, while the ratio cut (RC) [21] is chosen for measuring *inter-link* density.

Given a network $G = (V, E)$, suppose one division of G with k communities denoted as $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$. Let A be the adjacent matrix. Given one community C_m and $\overline{C_m} = \mathcal{C} - C_m$, $L(C_m, C_m)$ is defined as $\sum_{i \in C_m, j \in C_m} A_{ij}$ and $L(C_m, \overline{C_m})$ is defined as $\sum_{i \in C_m, j \in \overline{C_m}} A_{ij}$. Based on the above definitions, the two measures (KKM and RC) are formally defined as:

$$minimize \begin{cases} KKM = 2(|V| - k) - \sum_{i=1}^{k} \dfrac{L(C_i, C_i)}{|C_i|} \\ RC = \sum_{i=1}^{k} \dfrac{L(C_i, \overline{C_i})}{|C_i|} \end{cases} \quad (5)$$

From the above definitions, it can be observed that KKM can be considered as the sum of the *intra-link* densities, while RC can be considered as the sum of the *inter-link* densities. Thus, minimizing KKM and RC simultaneously can guarantee that the links within one community are dense while that between communities are sparse.

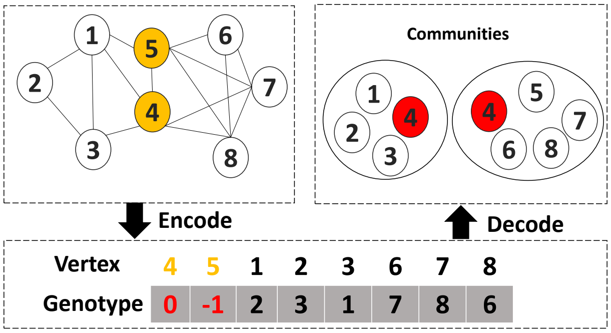
B. The Mixed-Representation Scheme

In this paper, the mixed representation suggested in [18] is used for fast encoding and decoding the network divisions. Specifically, the nodes in the network are classified into two groups: one group of candidate overlapping nodes and the other group of non-overlapping nodes. The length of this mixed-representation is equal to the size of the network, where the group of candidate overlapping nodes is denoted as the overlapping part of the individual and the group of non-overlapping nodes is denoted as the non-overlapping part of individual. The status for each candidate overlapping node can be '0' or '-1', where '0' indicates the corresponding node is an overlapping node while '-1' indicates the corresponding node is not an overlapping node. The status for each non-overlapping node can be the index of itself or its neighbor. For this mixed representation, it is easy and fast to decode the individual. Specifically, for decoding non-overlapping nodes, all nodes connected belong to one community. As for each candidate overlapping node, if its label is '0', then this node is assigned to all communities which the node connects to. Otherwise, if its label is '-1', then this node is assigned to the connected community with the maximum number of neighbors in this community. For the non-overlapping part, instead of the vector based individual representation scheme used in [18], the locus-based one is adopted in this strategy.

Fig. 1 presents an example to illustrate the mixed-representation. In this example, there are eight nodes in the network, where the group of candidate overlapping nodes is $\{4, 5\}$ and the group of non-overlapping nodes is $\{1, 2, 3, 6, 7, 8\}$. One individual ind is $\langle 0, -1, 2, 3, 1, 7, 8, 6\rangle$. When decoding the non-overlapping part, there are two local communities $\{1, 2, 3\}, \{6, 7, 8\}$. When decoding the overlapping part, the node 5 with status '-1' is assigned to the local community $\{6, 7, 8\}$ since it has three links which are larger than one link with the local community $\{1, 2, 3\}$. The node 4 with status '0' is assigned to the two local communities simultaneously. Thus, the final network divisions for ind are $\{1, 2, 3, 4\}, \{4, 5, 6, 7, 8\}$.

III. The Proposed Algorithm RMR-MOEA

In this section, we first present two individual length reduction strategies, and then give the general framework of the proposed algorithm RMR-MOEA.



Vertex	4	5	1	2	3	6	7	8
Genotype	0	-1	2	3	1	7	8	6

Fig. 1. An illustrative example for the mixed-representation.

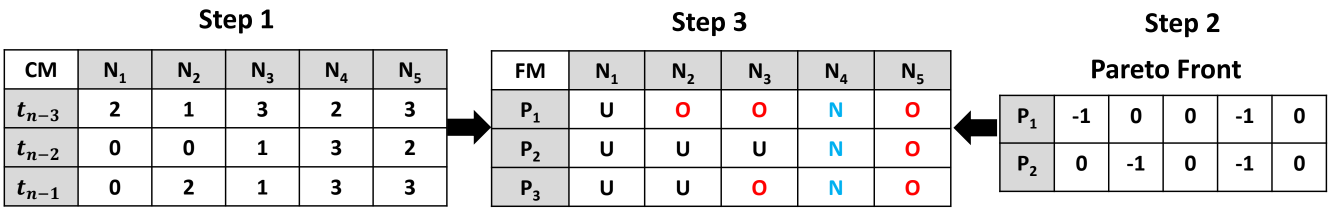
A. The Individual Length Reduction Strategies

In this subsection, we give two strategies for respectively shortening the individual length of overlapping part and non-overlapping part in the mixed representation.

1) The Historical Information Based Strategy for Fixing Overlapping Part: The main motivation of fixation strategy is based on the following observation. For candidate overlapping nodes, as the population evolves, they are gradually identified, so we can fix some candidate overlapping nodes that change stagnantly in previous generations. Hence, there are three main problems that need to be solved, that is, (1) which nodes should be fixed? (2) how many nodes should be fixed in the current population? and (3) what status of the fixation node is?

Firstly, we design a change matrix (denoted as CM) to record the historical information for each node in each population t_k, i.e., the total number of individuals in each generation whose corresponding node i has not changed in the previous generation, which is denoted as $TN_{i,k}$. In the second step, we use CM to calculate the number of individuals fixed in current population for each candidate overlapping node i (denoted as FN_i), and $FN_i = \lceil \sum_{k=1}^{K} TN_{i,k}/K \rceil$ (K is the number of rows in CM). Then, FN_i individuals are randomly selected from the current population (suppose $ind_{l1}, ind_{l2}, ..., ind_{lFN_i}$) for fixation of node i. Finally, we use the elite individuals of the current generation to vote the status of the fixation node i, denoted as S_i. To be specific, we choose more than half of the elite individuals with the same status as the fixed node's status. If the number of votes is the same, randomly choose the status of an overlapping node or the status of non-overlapping node. Then we get a fixation matrix (denoted as FM) by setting $FM_{lk,i}$ (k=1,2,3...FN_i) as S_i. Algorithm 1 presents the main procedure of the proposed strategy for fixing overlapping part.

Fig. 2 presents an example to illustrate the fixation operator, where CM preserves the historical information of the previous three generations and suppose there are three individuals in current population. In CM, the element $CM_{t_{n-3},N_3} = 3$ means that there are three individuals where the status of node N_3 has not changed in the t_{n-3} generation. For example, as for node N_3, we first calculate its fixation number, which is denoted as FN_3 and $FN_3 = \lceil 5/3 \rceil = 2$. Then, we randomly select two individuals from the current population, suppose P_1 and P_3 are selected to be fixed. Note that suppose



Step 1

CM	N_1	N_2	N_3	N_4	N_5
t_{n-3}	2	1	3	2	3
t_{n-2}	0	0	1	3	2
t_{n-1}	0	2	1	3	3

Step 3

FM	N_1	N_2	N_3	N_4	N_5
P_1	U	O	O	N	O
P_2	U	U	U	N	O
P_3	U	U	O	N	O

Step 2

Pareto Front

P_1	-1	0	0	-1	0
P_2	0	-1	0	-1	0

Fig. 2. An illustrative example for the historical information based strategy for fixing overlapping part, where the fixation matrix is denoted as FM, O represents overlapping node, N represents non-overlapping node and U represents uncertain node.

```
Algorithm 1: Evo-Fixation($CM, P$)
Input: $P$: the population; $CM$: the change matrix;
Output: $FM$: the fixation matrix;
$K \leftarrow$ the number of rows in $CM$;
$NodeNum \leftarrow$ the number of nodes in the network;
$Elites \leftarrow$ get the elite individuals of $P$ by non-dominated sorting;
for $i = 1$ to $NodeNum$ do
    $S_i \leftarrow$ get the status of node $i$ by voting in $Elites$;
end for
for $i = 1$ to $NodeNum$ do
    $FN_i \leftarrow \lceil \sum_{k=1}^{K} TN_{i,k}/K \rceil$;
    Randomly select $FN_i$ individuals (suppose $ind_{l1}, ind_{l2}, ..., ind_{lFN_i}$) in current population to fix;
    for $j = 1$ to $FN_i$ do
        $FM_{lj,i} \doteq S_i$ ;
    end for
end for
```

there are two elite individuals P_1 and P_2 in the current population and the status of node N_3 in P_1 and P_2 are the same (i.e. 0) so S_3 is denoted as O. Therefore, the value of N_3 in fixation matrix FM is fixed as overlapping node, i.e., $FM_{P_1,N_3} = O$ and $FM_{P_3,N_3} = O$. As for node N_4, the fixation number $FN_4 = \lceil 8/3 \rceil = 3$ and the status of node N_4 in the two elite individuals P_1 and P_2 are the same so S_4 is denoted as N, thus, the value of N_4 in fixation matrix FM is fixed as non-overlapping node, i.e., $FM_{P_1,N_4} = N$, $FM_{P_2,N_4} = N$ and $FM_{P_3,N_4} = N$. Similarly, for node N_5, $FM_{P_1,N_5} = O$, $FM_{P_2,N_5} = O$ and $FM_{P_3,N_5} = O$. From the above procedure, we can find that the number of fixed individual of overlapping part in the mixed representation can be increasing as the population evolves, so that the search space will be reduced.

2) The Local Communities Based Strategy For Reducing Non-Overlapping Part: The local community information is utilized to reduce the non-overlapping nodes since there often exist some nodes that belong to the same community in different population individuals. In other words, we get the same local community structure in elite individuals to merge as one new node to reduce the non-overlapping part. This idea is borrowed from [13]. To be specific, we first utilize non-dominant sorting to get the elite solutions which are denoted as P_1. Note that each individual represents a network division consisting of several communities.

The network reduction method is performed as follows. The individual $ind1$ with the largest number of communities in P_1 is selected. Then, for each community C_i in $ind1$, if all nodes in C_i are identified as in one community by the remaining elite individuals, then C_i is considered as a local community. Otherwise, we will regard the maximal subset of C_i which is belonged one community in remaining elite individuals as a local community. 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 nodes of each local community 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. Algorithm 2 presents the main procedure of the proposed strategy for reducing non-overlapping part.

```
Algorithm 2: Local-Merge($G, P$)
Input: $G$: the complex network; $P$: the population;
Output: $G_R$: the reduce network;
$P\_1 \leftarrow$ the non-dominated solutions in $P$;
$INum \leftarrow$ the number of individuals in $P\_1$;
$ind1 \leftarrow$ the individual in $P\_1$ with the largest number of communities;
$ComNum \leftarrow$ the number of the communities in $ind1$;
for $i = 1$ to $ComNum$ do
    $MerC_i \leftarrow$ the nodes set of $i$th community in $ind1$;
    for $j = 2$ to $INum$ do
        $ind_j\_C_k \leftarrow$ the $k$th community of $j$ individual, which has the most number of $ind_j\_C_k \cap MerC_j$;
        $MerC_i \leftarrow ind_j\_C_k \cap MerC_i$;
    end for
    $C_R \leftarrow$ Merge $MerC_i$ into one node;
end for
```

Fig. 3 presents an example to reduce the length of non-overlapping part. As shown in this figure, the local community $\{1,2,3\}$ is found since it appears in both $Ind1$ and $Ind2$. Then, the local community $\{1,2,3\}$ is considered as one node in the further evolution, thus the length of non-overlapping part is reduced from 7 to 3.

Combined with the above two strategies, it can be found that the individual length can be greatly reduced as the population evolves.

B. The General Framework

Based on the two individual length reduction strategies above, we present the general framework of RMR-MOEA, which is similar to MR-MOEA [18]. The RMR-MOEA consists of three steps. At the first step, the network is reduced by using the local topology structure of the network according to a pre-reduction method [13]. The candidate overlapping nodes

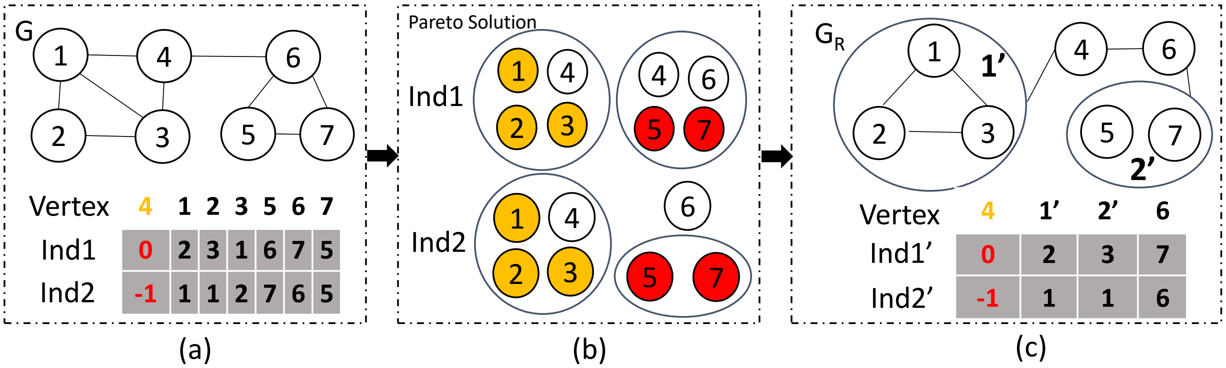


Fig. 3. An example of the reduced strategy for non-overlapping nodes. The local communities information is utilized to shorten the size of non-overlapping nodes by merging each local community as one node.

O are found according to the method proposed in [18], and the size of O is denoted as os.

At the second step, a population with pop individuals is initialized based on the mixed-representation. Specifically, each individual has two parts, the overlapping part, $X_i[1 : os]$, represents candidate overlapping nodes, randomly assigned with -1 or 0, which is binary encoding. The non-overlapping part, $X_i[os + 1 : end]$, represents non-overlapping nodes, randomly assigned with the index of its neighbors or itself. The velocity is assigned with the vector **0**. The reference point z^* is initialized by using the best values of two optimized objectives KKM and RC (see Eq. (5)) 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. We initialize the CM assigned with the 0 and the FM with the status of U (uncertain node).

At the third step, during the evolution, we firstly randomly select an individual from N_j as the $Gbest_i$ of X_i. Then the new velocity V_i will be computed by $Gbest_i$ and $Pbest_i$. The new position $child_i[1 : os]$ are generated by V_i, utilizing the particle swarm optimization (PSO) algorithm operator [22]. For reducing the search space, we utilize the algorithm1 to fix the position. The new position $child_i[os + 1 : end]$ are generated by Genetic algorithm(GA) operators like crossover and mutation [21]. If the Tchebycheff value of $chlid_i$ is better than an individual in N_j, then replace the individual and update reference point z^*. We will record the change information in CM and if the $child_i$ dominates $Pbest_i$, we will update the $Pbest_i$. When i-th generation satisfying that $i \mid (\lceil (maxgen+1)/(T+1) \rceil) == 0$, we run the suggested two individual length reduction strategies for current population, where T is a parameter for controlling the number of strategies executed and naturally the number of rows in CM is set T. As the length of population may change, we should update the P and z^*. Algorithm 3 gives the main procedure of RMR-MOEA.

Algorithm 3: General Framework of RMR-MOEA

```
Input: $G$: the complex network; $gene$: the number of generations;
    $pop$: the size of population; $\{\lambda_1, \lambda_2, \ldots, \lambda_{pop}\}$: the set of
    weight vectors; $ns$: the size of neighbours; $p_c$: crossover
    probability, $p_m$: mutation probability; $T$: the times of running
    the individual length reduction strategies; $c1, c2$:the learning
    factors; $w$:the inertia weight;
Output: Optimal solutions
Step1: the candidate overlapping nodes and subgraph finding
$O \leftarrow$ utilize candidate overlapping nodes finding method
    in [18] and get the size of $O$ denoted as $os$;
$G_R \leftarrow$ utilize the pre-reduction method in [13] for $G$;
Step2: initialization
Position initialization:$P = \{X_1, X_2, \ldots, X_{pop}\}$;
Pbest initialization:$Pbest = \{X_1, X_2, \ldots, X_{pop}\}$
Velocity initialization:$V = \{V_1, V_2, \ldots, V_{pop}\}$
Change matrix initialization: $CM = \{0, 0, 0...\}$;
Fixation matrix initialization: $FM = \{U, U, U...\}$;
Initialize reference point $z^*$;
Step3: population evolution
$N = \{N_1, N_2, \ldots, N_{pop}\} \leftarrow$ obtain the neighbors of each
    individual by computing Euclidean distance based on the set of
    weight vectors;
for $t = 1$ to $gene$ do
    if $i \mid (\lceil (maxgen + 1)/(T + 1) \rceil) == 0$ then
        $[FM, P_R] \leftarrow$ execute historical information based strategy
            for fixing overlapping part (Evo-Fixation($CM$, $P$));
        $[G_R, P_R] \leftarrow$ execute the local communities based strategy
            for reducing non-overlapping part (Local-Merge($G$, $P$));
        Update $P$ and reference point $z^*$ based on $G_R$;
    end if
    for $i = 1$ to $pop$ do
        Randomly select one individual from $N_i$ as $Gbest_i$;
        $V_i \leftarrow$ compute velocity by $Pbest_i$ and $Gbest_i$;
        $child_i[1 : os] \leftarrow$ generate child by PSO operators and
            utilize the $FM$ to fix;
        $child_i[os + 1 : end] \leftarrow$ generate child by GA operators;
        Compute objective function;
        Update $P$;
        Record position change in $CM$;
        Update reference point $z^*$ and $Pbest_i$;
    end for
end for
```

IV. Experimental Results

In this section, we first give experimental settings, including comparison algorithms, real-world networks and evaluation criterion. Then, we present the comparison results between RMR-MOEA and baselines.

A. Experimental Settings

1) Comparison Algorithms: In this paper, five representative algorithms are chosen to compare with the proposed RMR-MOEA. Specifically, RMR-MOEA is compared with four MOEA-based overlapping community detection algorithms (namely IMOQPSO [16], MCMOEA [17], LG-MOEA [20] and MR-MOEA [18]) and one non-MOEA-based algorithm LMD [23]. For each baseline algorithm, we use the code provided by the author and adopt its parameters suggested in their paper.

For a fair comparison, in the four MOEA-based algorithms, the population size PS is all set to 100 and the maximum number of generations $gene$ is set to 100. The experimental results for all algorithms are obtained by averaging over 15 independent runs. All the experiments are carried out on computers with Intel Core i7-8700K 3.70 GHz CPU, 32-GB RAM and Windows 10 operating system.

2) Real-World Networks: We adopt 10 popular real-world networks with different characteristics to evaluate the performance of comparison algorithms. These networks are Zacharys karate club [24], Dolphin social network [25], American college football [26], Books about US politics [26], Scientist collaboration network [25], Yeast PPI dataset [27], Blogs network [18], CA-HepTh1 [28], PGP [29], CA-HepTh2 [28]. The characteristics of these networks are given in Table I. Note that karate, dolphin, football, polbooks, are networks with ground truth community structure.

TABLE I
10 REAL-WORLD NETWORKS WITH DIFFERENT CHARACTERISTICS.

Networks	Nodes	Edges	Ave. Degree	Real Clusters
karate	34	78	4.59	2
dolphin	62	159	5.13	2
football	115	613	10.66	12
polbook	105	441	8.4	3
netscience	1,589	2,742	3.45	Unknown
PPI	2,456	6,265	5.26	Unknown
blogs	3,984	6,803	3.41	Unknown
ca-HepTh1	9,877	25,998	5.26	Unknown
PGP	10,680	24,316	4.55	Unknown
ca-HepTh2	12,008	118,521	19.74	Unknown

3) Evaluation Criterion: In this paper, we adopt the generalized normalized mutual ($gNMI$) [30] and the extended modularity Q_{ov} to evaluate the quality of overlapping communities detected.

The $gNMI$ is used to measure the similarity between detected community partition and the real community partition. Specifically, $gNMI$ can be utilized and defined as follows.

$$gNMI(C_t, C_d) = \frac{-2\sum_{i=1}^{k1}\sum_{j=1}^{k2} M_{ij}\log\frac{M_{ij}N}{M_{i*}M_{*j}}}{\sum_{i=1}^{k1} M_{i*}\log(\frac{M_{i*}}{N}) + \sum_{j=1}^{k2} M_{*j}\log(\frac{M_{*j}}{N})} \tag{6}$$

where C_t represents the true community division, C_d is a community division that is to be evaluated detected by an algorithm. $k1$ and $k2$ denote the numbers of communities in C_t and C_d respectively. M represents the confusion matrix. The number of rows and columns of M are $k1$ and $k2$ respectively. M_{ij} is the number of shared nodes in the i-th community of C_t and the j-th community of C_d. Moreover, M_{i*} is the number summed by M in row i, and M_{*j} is the number summed by M in column j. The number of nodes in the network is denoted as N. When $gNMI$ is 1, it represents that the reality of network community division is found completely by the algorithm. The larger the $gNMI$ is, the better the performance of the algorithm is.

The another criterion Q_{ov} is measured for the detected overlapping communities when the truth community division is unknown. Specifically, Q_{ov} can be calculated by the following formula.

$$Q_{ov} = \frac{1}{2m}\sum_{k}\sum_{i,j\in C_k}\frac{1}{P_iP_j}(A_{ij} - \frac{d_id_j}{2m}) \tag{7}$$

where the number of edges in network is m. C_k represents the k-th community in division. d_i and d_j represent the node degree of node i and j respectively. The number of communities i belongs to is P_i. A_{ij} indicates the value of the adjacency matrix A in the i-th row and j-th column. Note that the larger Q_{ov} is, the better the quality of the overlapping communities division is.

B. The Comparison Results between RMR-MOEA and Baselines

In the following, we first give the comparison results in terms of Q_{ov} on 10 real datasets, and then present the comparison results in terms of $gNMI$ on the four real datasets with ground truth community structure. Finally, the running time between RMR-MOEA and baseline RMR-MOEA is also compared. Note that, the best value of Q_{ov} and $gNMI$ in the obtained non-dominated solution of the MOEA-based algorithms is used for comparison since this way has been widely adopted in existing MOEA-based community detection algorithm for comparing the performance [17], [18].

1) Experimental Results in Terms of Q_{ov}: Table II shows the Q_{ov} values of the proposed algorithm RMR-MOEA and the other five community detection algorithms on the 10 real-world networks. We adopt the Wilcoxon rank sum test at a significance level of 0.05 to evaluate the statistical difference of the performance of comparison algorithms, where the symbols '+', '-' and '≈' indicate that the result is significantly better, significantly worse and statistically similar to that obtained by RMR-MOEA, respectively. From this table, we can find that the proposed RMR-MOEA achieves the best performance on most of the real-world networks in terms of Q_{ov}. The baseline MR-MOEA or LG-MOEA achieve the second best performance on most of the real-world networks.

2) Experimental Results in Terms of $gNMI$: In order to further show the performance of the proposed algorithm, we also adopt $gNMI$ as the performance metric. However, $gNMI$ can only be used in the datasets with the ground truth community structure. In our experiments, there are four datasets with real community structure. Table III shows the $gNMI$ values of RMR-MOEA and the other five community detection algorithms. From this table, it can be found that the proposed RMR-MOEA achieves the best performance on most

TABLE II

THE COMPARISON RESULTS OF Q_{ov} ON THE 10 REAL-WORLD NETWORKS, WHERE SYMBOLS '+', '-' AND '≈' INDICATE THAT THE PERFORMANCE IS SIGNIFICANTLY BETTER, SIGNIFICANTLY WORSE AND STATISTICALLY SIMILAR TO THAT OF RMR-MOEA, RESPECTIVELY. NOTE THAT '/' MEANS THAT Q_{ov} VALUE IS NOT PROVIDED HERE SINCE THE RESULT CANNOT BE OBTAINED WITHIN 14 HOURS FOR ONE RUN.

Network	Metric	RMR-MOEA	MR-MOEA	IMOQPSO	MCMOEA	LMD	LGMOEA
karate	Q_{ov}_max	0.223	**0.230**	0.184	0.212	0.216	0.210
	Q_{ov}_avg	0.219(10.8s)	$\mathbf{0.221}^{\approx}$(25.3s)	0.196^{-}	0.209^{-}	0.211^{-}	0.208^{-}
	Std	0.001	0.004	0.006	0.006	0.003	0.003
dolphin	Q_{ov}_max	**0.274**	0.271	0.153	0.213	0.261	0.253
	Q_{ov}_avg	**0.268**(17.1s)	0.261^{-}(56.1s)	0.132^{-}	0.199^{-}	0.252^{-}	0.232^{-}
	Std	0.003	0.006	0.008	0.009	0.017	0.015
football	Q_{ov}_max	**0.303**	0.302	0.235	0.279	0.291	0.300
	Q_{ov}_avg	**0.298**(24.9s)	$0.297^{\approx}$(90.7s)	0.229^{-}	0.274^{-}	0.284^{-}	$0.296^{\approx}$
	Std	0.001	0.004	0.005	0.005	0.007	0.004
polbook	Q_{ov}_max	**0.269**	0.265	0.172	0.239	0.259	0.267
	Q_{ov}_avg	**0.266**(22.3s)	0.262^{-}(78.3s)	0.165^{-}	0.215^{-}	0.250^{-}	0.245^{-}
	Std	0.001	0.002	0.012	0.013	0.009	0.015
netscience	Q_{ov}_max	**0.473**	0.470	0.289	0.453	0.397	0.375
	Q_{ov}_avg	**0.470**(866s)	0.465^{-}(2030s)	0.247^{-}	0.449^{-}	0.395^{-}	0.372^{-}
	Std	0.002	0.002	0.033	0.002	0.015	0.014
PPI	Q_{ov}_max	0.291	**0.310**	0.256	0.207	0.277	0.293
	Q_{ov}_avg	0.280(3243s)	$\mathbf{0.301}^{+}$(7773s)	0.252^{-}	0.199^{-}	0.274^{-}	$0.283^{\approx}$
	Std	0.003	0.002	0.030	0.004	0.002	0.003
blogs	Q_{ov}_max	0.372	0.394	0.340	0.156	0.322	**0.395**
	Q_{ov}_avg	0.369(5605s)	0.385^{+}(13970s)	0.336^{-}	0.144^{-}	0.317^{-}	$\mathbf{0.393}^{+}$
	Std	0.002	0.008	0.006	0.024	0.014	0.015
ca-HepTh1	Q_{ov}_max	0.271	/	0.245	0.107	0.215	**0.291**
	Q_{ov}_avg	0.269(27378s)	/	0.104^{-}	0.104^{-}	0.213^{-}	$\mathbf{0.288}^{+}$
	Std	0.001	/	0.003	0.002	0.001	0.001
PGP	Q_{ov}_max	0.368	/	/	/	0.342	**0.388**
	Q_{ov}_avg	0.365(26193s)	/	/	/	0.341^{-}	$\mathbf{0.385}^{+}$
	Std	0.002	/	/	/	0.001	0.015
ca-HepTh2	Q_{ov}_max	**0.216**	/	/	/	/	/
	Q_{ov}_avg	**0.213**(45573s)	/	/	/	/	/
	Std	0.001	/	/	/	/	/
$+/-/\approx$		—	2/6/2	0/10/0	0/10/0	0/10/0	3/6/1

of the real-world networks in terms of $gNMI$. The baseline MR-MOEA achieves the second best performance.

3) Experimental Results in Terms of Running Time: For large-scale community detection problems, running time is also a very important criterion to evaluate the performance. In order to show the efficiency of RMR-MOEA, we compare RMR-MOEA with MR-MOEA in terms of running time. Table II also shows the average running time values of RMR-MOEA and MR-MOEA, where the values are in the parentheses behind Q_{ov}_avg value. As can be found from this table, RMR-MOEA takes much less time than MR-MOEA on all networks, especially for large-scale networks. For example, for network ca-HepTh1, MR-MOEA can not obtain the result within 14h for only one run while RMR-MOEA can get the final result with 7.6h (27,378s).

Based on the empirical results shown in Tables II and III, we can conclude that the proposed RMR-MOEA algorithm holds a competitive performance in terms of both detection performance (Q_{ov} and $gNMI$) and running time on real-world networks. The better performance of RMR-MOEA is attributed to the proposed individual length reduction strategies, which can be used to greatly reduce the search space.

V. CONCLUSIONS

In this paper, we proposed a reduced mixed representation based multi-objective evolutionary algorithm named RMR-MOEA for overlapping community detection on large-scale complex networks, where the length of the individual is recursively shortened as the evolution proceeds. To be specific, we adopt a mixed representation suggested in MR-MOEA for fast encoding and decoding the individual, which consists of potential overlapping nodes part and non-overlapping nodes part. Then, in each length reduction of individual, the historical information was used to fix potential overlapping nodes, while the local communities existed in elite individuals were used to reduce non-overlapping nodes. Based on these two length reduction strategies, the search space of RMR-MOEA can be greatly reduced. Finally, we compared RMR-MOEA with five representative baselines on 10 real-world complex networks and the experimental results demonstrate the effectiveness of the proposed RMR-MOEA in terms of both detection performance and running times, especially on large-scale networks.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (61976001, 61876184, 62076001), and also supported by the Natural Science Foundation of Anhui Province (2008085QF309, 1908085MF219).

TABLE III

THE COMPARISON RESULTS OF $gNMI$ ON THE FOUR REAL-WORLD NETWORKS WITH GROUND TRUTH, WHERE SYMBOLS '+', '-' AND '≈' INDICATE THAT THE PERFORMANCE IS SIGNIFICANTLY BETTER, SIGNIFICANTLY WORSE AND STATISTICALLY SIMILAR TO THAT OF RMR-MOEA, RESPECTIVELY.

Network	Metric	RMR-MOEA	MR-MOEA	IMOQPSO	MCMOEA	LMD	LGMOEA
karate	$gNMI\_max$	**1**	**1**	0.634	0.918	0.422	**1**
	$gNMI\_avg$	**1**	$\mathbf{1}^{\approx}$	0.491^{-}	0.839^{-}	0.365^{-}	0.836^{-}
	Std	0	0.092	0.061	0.030	0.081	0.065
dolphin	$gNMI\_max$	**1**	**1**	0.221	0.176	0.611	0.888
	$gNMI\_avg$	**0.971**	0.878^{-}	0.137^{-}	0.198^{-}	0.287^{-}	0.5316^{-}
	Std	0.028	0.024	0.017	0.026	0.009	0.312
football	$gNMI\_max$	**0.828**	0.821	0.226	0.300	0.781	0.801
	$gNMI\_avg$	**0.801**	0.784^{-}	0.209^{-}	0.274^{-}	0.744^{-}	0.779^{-}
	Std	0.026	0.028	0.044	0.013	0.029	0.031
polbook	$gNMI\_max$	**0.500**	**0.500**	0.296	0.192	0.233	0.395
	$gNMI\_avg$	**0.500**	0.419^{-}	0.219^{-}	0.168^{-}	0.199^{-}	0.331^{-}
	Std	0.000	0.004	0.042	0.046	0.032	0.063
$+/-/\approx$		—	0/3/1	0/4/0	0/4/0	0/4/0	0/4/0

REFERENCES

[1] R. Pastor-Satorras and A. Vespignani, "Evolution and structure of the internet: A statistical physics approach," 2007.

[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," *Bioinform*, vol. 30, no. 10, pp. 1343–1352, 2014.

[3] S. Wasserman and K. Faust, *Social network analysis - methods and applications*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 2007, vol. 8.

[4] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[5] C. Pizzuti, "Evolutionary computation for community detection in networks: A review," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 3, pp. 464–483, 2018.

[6] ——, "A multiobjective genetic algorithm to find communities in complex networks," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 3, pp. 418–430, 2012.

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

[8] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 82–97, 2013.

[9] D. Chen, F. Zou, R. Lu, L. Yu, Z. Li, and J. Wang, "Multi-objective optimization of community detection using discrete teaching–learning-based optimization with decomposition," *Information Sciences*, vol. 369, pp. 402–418, 2016.

[10] K. R. Žalik and B. Žalik, "Multi-objective evolutionary algorithm using problem-specific genetic operators for community detection in networks," *Neural Computing and Applications*, vol. 30, no. 9, pp. 2907–2920, 2018.

[11] F. Cheng, T. Cui, Y. Su, Y. Niu, and X. Zhang, "A local information based multi-objective evolutionary algorithm for community detection in complex networks," *Applied Soft Computing*, vol. 69, pp. 357–367, 2018.

[12] F. Zou, D. Chen, D.-S. Huang, R. Lu, and X. Wang, "Inverse modelling-based multi-objective evolutionary algorithm with decomposition for community detection in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 513, pp. 662–674, 2019.

[13] X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, and Y. Jin, "A network reduction-based multiobjective evolutionary algorithm for community detection in large-scale complex networks," *IEEE Transactions on Cybernetics*, vol. 50, no. 2, pp. 703–716, 2020.

[14] J. Liu, W. Zhong, H. A. Abbass, and D. G. Green, "Separated and overlapping community detection in complex networks using multiobjective evolutionary algorithms," in *IEEE Congress on Evolutionary Computation*. IEEE, 2010, pp. 1–7.

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

[16] Y. Li, Y. Wang, J. Chen, L. Jiao, and R. Shang, "Overlapping community detection through an improved multi-objective quantum-behaved particle swarm optimization," *Journal of Heuristics*, vol. 21, no. 4, pp. 549–575, 2015.

[17] X. Wen, W.-N. Chen, Y. Lin, T. Gu, H. Zhang, Y. Li, Y. Yin, and J. Zhang, "A maximal clique based multiobjective evolutionary algorithm for overlapping community detection," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 3, pp. 363–377, 2016.

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

[19] Y. Tian, S. Yang, and X. Zhang, "An evolutionary multiobjective optimization based fuzzy method for overlapping community detection," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 11, pp. 2841–2855, 2020.

[20] H. Ma, H. Yang, K. Zhou, L. Zhang, and X. Zhang, "A local-to-global scheme-based multi-objective evolutionary algorithm for overlapping community detection on large-scale complex networks," *Neural Computing and Applications*, pp. 1–15, 2020.

[21] M. Gong, L. Ma, Q. Zhang, and L. Jiao, "Community detection in networks by using multiobjective evolutionary algorithm with decomposition," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 15, pp. 4050–4060, 2012.

[22] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Transactions on Evolutionary Computation*, 2014.

[23] D. Jin, B. Gabrys, and J. Dang, "Combined node and link partitions method for finding overlapping communities in complex networks," *Scientific Reports*, vol. 5, p. 8600, 2015.

[24] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[25] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl_2, pp. S186–S188, 2003.

[26] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[27] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang *et al.*, "Topological structure analysis of the protein–protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[28] J. Leskovec and A. Krevl, "Snap datasets: Stanford large network dataset collection," 2014.

[29] X. Guardiola, R. Guimera, A. Arenas, A. Diaz-Guilera, D. Streib, and L. Amaral, "Macro-and micro-structure of trust networks," *ArXiv Preprint Cond-Mat/0206240*, 2002.

[30] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.