ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20150317103835

# AN EFFICIENT MULTI-OBJECTIVE COMMUNITY DETECTION ALGORITHM IN COMPLEX NETWORKS

*Kun Deng, Jian-Pei Zhang, Jing Yang*

Original scientific paper
Community detection in complex networks is often regarded as the problem of single-objective optimization and it is hard for single-objective optimization to identify potential community structure of meaningfulness. Thus, algorithm of multi-objective optimization is applied to the field of community detection. However, multi-objective community detection algorithm is prone to local optimization and weak diversity of the set of Pareto-optimal solutions. In view of this, based on the framework of NSGAII, a multi-objective community detection algorithm, named I-NSGAII, is proposed in this paper. This algorithm is able to optimize simultaneously the two conflicting objective functions evaluating the density of intra-community connections and the sparsity of inter-community connections, and obtain the set of Pareto optimal solutions having diverse hierarchal community structures; it also proposes diversity evolutionary strategy enabling the algorithm to expand searching space and thus avoids local optimization of the set of Pareto-optimal solutions. In addition, to improve algorithm's searching ability, I-NSGAII algorithm adopts the strategies of locus-based adjacency representation, unified label, one-way crossover and local mutation. Tests on synthetic and real-world networks and comparisons with many state-of-the-art algorithms verify the validity and feasibility of I-NSGAII.

***Keywords:*** *complex networks; diversity evolutionary strategy; multi-objective community detection; NSGAII*

## Učinkoviti višekriterijski algoritam za detekciju zajednice u složenim mrežama

Izvorni znanstveni članak
Detekcija zajednice u složenim mrežama često se smatra problemom jednokriterijske optimizacije, a teško je jednokriterijskom optimizacijom identificirati moguću strukturu zajednice punu značenja. Stoga je algoritam višekriterijske optimizacije primijenjen na područje detekcije zajednice. Međutim, algoritam višekriterijske detekcije zajednice sklon je lokalnoj optimizaciji i slaboj raznovrsnosti niza Pareto-optimalnih rješenja. Imajući to u vidu, u ovom se radu predlaže višekriterijski algoritam za detekciju zajednice, nazvan I-NSGAII, zasnovan na sustavu NSGAII. Taj algoritam može simultano optimizirati dvije suprotstavljene kriterijske funkcije procjenjujući gustoću veza unutar zajednice i nedostatak veza između zajednica te dobiti niz Pareto optimalnih rješenja koja imaju strukturu zajednice različitih hijerarhija; on također predlaže razvojnu strategiju raznolikosti (diverziteta), omogućujući algoritmu proširenje područja pretraživanja te tako izbjegava lokalnu optimizaciju niza Pareto-optimalnih rješenja. Uz to, kako bi se poboljšala mogućnost pretraživanja algoritma, I-NSGAII algoritam usvaja strategije predstavljanja susjedstva prema mjestu (locus-based adjacency representation), jedinstvenog naziva, crossovera u jednom smjeru i lokalne mutacije. Ispitivanja na sintetičkim i mrežama stvarnog svijeta te usporedbe s mnogim state-of-the-art algoritmima potvrđuju validnost i izvedivost I-NSGAII-a.

***Ključne riječi:*** *NSGAII; razvojna strategija diverziteta; složene mreže; višekriterijska detekcija zajednice*

## 1 Introduction

Many complex systems in real world can be represented as complex networks such as social networks, biological networks, technology networks and Web networks etc. Analyses of these networks discovered their statistical properties such as "small-world effect" [1] and "power laws" in the link distribution [2]. And another property which receives particular attention is the property of community structure. The property of "community structure" refers to the characteristics of dense intra-community connections and sparse inter-community connections which largely exist in complex networks.

The content of community structure varies according to different application fields. For example, it refers to people sharing similar characteristics in social networks and the Bio-Modules with similar functions in biological networks etc. The purpose of community detection in complex networks is to reveal community structure inherent in those complex networks. Its study is of great significance both theoretically and practically, and has attracted many researcher of various disciplines and been applied to numerous fields such as terrorist organization recognition, biological network analysis, Web community mining, link prediction etc.

In recent years, optimization algorithm based on modularity (Q) [3] function emerges in an endless stream, such as Fast Newman (FN) algorithm [4] and Fast Unfolding Algorithm (FUA) [5] etc. However, it analyses community structure only from single perspective while optimizing single-objective function and conducting community detection; and the corresponding community structure of the optimal objective function value obtained by it is usually the inappropriate community structure [6].

Different from single-objective community detection, multi-objective community detection is based on Pareto optimality theory. Instead of finding a single optimal solution, it surveys the community structure in complex networks from different angles and identifies a set of Pareto-optimal solutions with each of these solutions corresponding to a trade-off between multiple objective functions, thus it can avoid the risk that single-objective community detection may only be suitable to one kind of networks. In addition, multi-objective community detection algorithm is capable of combining various evaluation criteria of community quality, thus it is conducive to the discovery of meaningful community structure in complex networks.

In order to solve the limitation of single-objective community detection, a multi-objective genetic algorithm for community detection, named MOGANET [7], was presented by Pizzuti in 2012. This algorithm, through simultaneously optimizing the two objective functions of Community Score and Community Fitness, gets a set of Pareto-optimal solutions with each of these solutions corresponding to a trade-off between these two objective functions. Although multi-objective community detection is capable of surveying community structure from different angles and uncovering potential community structure of meaningfulness in complex networks, the exertion of these advantages needs the assurance of its convergence efficiency together with the diversity of Pareto optimal solutions. Nevertheless, during its processes of evolution and selection, some elite solutions will dominate the whole space of solutions and prevent the entering of new ones, which could lead to the defects of local optimization and decrease in diversity of the set of Pareto-optimal solutions. Thus, strengthening the diversity of the set of Pareto-optimal solutions while assuring the convergence efficiency of multi-objective community detection has become non-negligible in the field of multi-objective community detection.

In view of those mentioned above, in this paper, based on the framework of NSGAII (no dominated sorting genetic algorithm II) [8], the algorithm of I-NSGAII is proposed and is applied to the field of multi-objective community detection. Through simultaneously optimizing two conflicting objective functions, it gets a set of Pareto optimal solutions with each of these solutions corresponding to a community structure and to a trade-off between these two objective functions. It also proposes diversity evolutionary strategy with the view of strengthening the diversity and avoiding the local optimization of the set of Pareto-optimal solutions. Besides, I-NSGAII presents the integrated application of strategies of locus-based adjacency representation [7], unified label, one-way crossover [9] and local mutation [10] to improve algorithm's searching ability. Analysis of experiment verifies that I-NSGAII is able to strengthen the diversity of the set of Pareto-optimal solutions while searching it fast, thus it can avoid the local optimization of the set of Pareto optimal solutions and is easy to uncover the community structure of meaningfulness.

## 2 Related background and motivations

### 2.1 Multi-objective optimization

**Definition 1 (Pareto dominated)**. Suppose $P$ is a set of feasible solutions, the size of it is $n$, and the number of property for each individual within it is $m$, $f_k$ is the criterion function for each property ($k$=1, 2, …, $m$). Given $x, y \in P$ , $y$ is said to dominate $x$ and $y$ is called Pareto dominated, denoted as $y \succ x$ , if and only if

$$\begin{aligned}&(\forall i \in \{1,2,...,m\}: f_i(y) \le f_i(x))\\ &\wedge(\exists k \in \{1,2,...,m\}: f_k(y) < f_k(x))\end{aligned} \tag{1}$$

**Definition 2 (Pareto optimal solution)**. Given $y \in P$ , $y$ is called the Pareto optimal solution, if and only if

$$\neg\exists x \in P : x \succ y \tag{2}$$

**Definition 3 (set of Pareto optimal solutions)**. The set $P_s$ composed by the Pareto optimal solutions is called the set of Pareto-optimal solutions and is defined as

$$P_s = \{y \mid \neg\exists x \in P : x \succ y\} \tag{3}$$

**Definition 4 (Pareto front)**. The solutions within the set $P_s$ correspond to objective function values which compose set $P_F$. The $P_F$ is called the Pareto front and is defined as

$$P_F = \{F(x) = (f_1(x), f_2(x), ..., f_m(x)) \mid x \in P_s\} \tag{4}$$

### 2.2 Framework of NSGAII

NSGAII transforms multiple evaluation criteria to single fitness assignment using the Pareto dominated relationship, and divides the individuals in the population into the different Pareto fronts by Pareto dominated sorting. After Pareto fronts are created, their interior members are sorted according to the crowding distance assignment.

The basic operation of NSGAII is as follows: in each generation of genetic operation, $N$ new individuals are generated ($N$ denotes the size of population). And then $N$ optimal individuals are selected from the new individuals and parent individuals as the parent individuals of the next generation genetic operation. The above-mentioned operation is repeatedly implemented until a huge elite set can be kept from generation to generation.

Since NSGAII can avoid the loss of elite individuals, the framework of NSGAII is taken as the algorithm framework of I-NSGAII.

### 2.3 Related work

In recent years, various kinds of approaches to detect community structure have been proposed. GN algorithm [11] was proposed by Newman. In GN, the edge betweenness of each edge is first presented, which is defined as the number of the shortest paths from nodes of one community to nodes of the other communities. And these paths pass through the inter-community edges. And then network is split by repeatedly computing edge betweenness and deleting the edge with the maximum edge betweenness until the dendrogram concerning community structure is obtained. In order to get the optimal community structure from the dendrogram, Newman defined modularity (Q). Thereafter, optimal algorithm based on modularity has become the mainstream algorithm in the field of community detection, such as Newman's FN algorithm. It first initializes every node within network to a single community; then during each process of iteration, it chooses and combines the two communities able to increase/decrease the modularity to the highest degree; and finally it combines the whole network as one community. A dendrogram about the community structure is established through the above-mentioned process. At last, the community structure with maximal $Q$ function value is used as the final detection result. Research shows [12] that the community detection method to precisely calculate the maximum modularity is NP-complete problem. Thus, as an effective method of solving NP-complete problem, genetic algorithm has been widely applied to the field of community detection. For example, GANET algorithm [13] proposes community score to measure the quality of community detection, and introduces the graph-based encoding strategy to community detection; then it completes the task of community detection based on the three traditional genetic strategies of uniform crossover, random mutation and elite selection.

The above mentioned algorithms analyse community structure from single perspective. The corresponding community structures of the optimal objective function values obtained by those algorithms are usually inappropriate [6]. In view of this, multi-objective community detection algorithm was proposed. For example, MOGANET [7] algorithm, presented by Pizzuti in 2012, introduced to it two objective functions. First, it adopts community score (*CS*) to measure the density of intra-community connections—the higher the *CS* value, the denser the intra-community connections; second, it uses community fitness (*CF*) to evaluate the sparsity of inter-community connections—when the value of *CF* reaches the maximum, the number of intra-community connections is the maximum and number of inter-community connections is the minimum. Under the framework of genetic algorithm, this algorithm, by optimizing the two functions of *CS* and *CF*, obtains a set of Pareto optimal solutions with each of these solutions corresponds to a trade-off between these two objective functions, and thus completes the task of community detection.

In 2012, Gong et al proposed Multi-objective Evolutionary Algorithm based on Decomposition (MOEA/D) [14]. Within the framework of genetic algorithm, this method, through simultaneously optimizing Negative Ratio Association (*NRA*) function which evaluates the density of intra-community connections and Ratio Cut (*RC*) function which evaluates the sparsity of inter-community connections, gets a set of Pareto optimal solutions with each of these solutions corresponds to a trade-off between *NRA* and *RC* functions, and thus achieves the purpose of community detection.

In 2013, Huang et al designed Multi-objective Community Detection based on particle swarm optimization (MOCD-PSO) [15]. This algorithm simultaneously optimizes modularity function evaluating the density of intra-community connections, MinMaxCut (MMC) function evaluating the sparsity of inter-community connections, and Silhouette function evaluating the similarity of nodes within community. And then it obtains a set of Pareto optimal solutions which corresponds to different community structures and completes the task of community detection.

Despite the fact that multi-objective community detection is able to survey structure property in complex networks from multiple angles and to identify potential community structure of meaningfulness in complex networks, the exertion of these advantages needs to consider the diversity of the set of Pareto optimal solutions while ensuring algorithm's convergence efficiency.

Consequently, the algorithm of I-NSGAII is designed in this paper which could effectively enhance the diversity of the set of Pareto optimal solutions. This method is able to search the Pareto optimal solutions fast and at the same time enlarge the searching space of Pareto optimal solutions, thus it can avoid the local optimization of the set of Pareto optimal solutions.

## 2.4 Our Motivations

### 2.4.1 Motivations for selecting objective functions

When conducting multi-objective detection, the selection of objective functions should pay attention to community structure in complex networks from multiple perspectives. And to improve the universality of algorithm, the input parameters of function need to be reduced to the most degree. Thus, members tightly-knit function (MT) evaluating the density of intra-community connections, is proposed in this paper and combined with average conductance function (AC) evaluating the sparsity of inter-community connections. And I-NSGAII algorithm takes these two functions as the objective functions. These two functions differ with each other in their observing perspectives and there is no need to input any parameters. Consequently, these two objective functions adopted in this paper are suitable to community detection.

### 2.4.2 Motivations for presenting diversity evolutionary strategy

During multi-objective community detection's processes of evolution and selection, some elite solutions will dominate the whole space of solutions and prevent the entering of new ones, which could lead to the defects of local optimization and decrease of diversity of the set of Pareto-optimal solutions. Therefore, this paper presents that diversity evolutionary strategy should be added to the framework of NSGAII. Under the condition that the number of individual is guaranteed to be stable, this strategy evenly selects several individuals from every Pareto front, and then conducts on those selected individuals the operation of overall mutation for neighbour nodes which is able to expand the space of solutions. Consequently, it increases the diversity of Pareto optimal solutions, enlarges the searching space of the set of Pareto optimal solutions, and avoids the local optimization of the set of Pareto optimal solutions.

### 2.4.3 Motivations for selecting evolutionary strategy

Within multi-objective community detection, excellent evolutionary strategy is indispensable to find the set of Pareto optimal solutions fast. Thus, to improve the searching ability of algorithm, this paper proposes the integrated application of the strategies of locus-based adjacency representation, one-way crossover and local mutation; at the same time, it also presents the strategy of unified label to solve the problem that the strategy of locus-based adjacency representation is hard to calculate the value of objective function.

## 3 The proposed I-NSGAII for community detection

### 3.1 Objective functions

Based on the intuitive understanding of community structure that within community every node should be in the same community with its most neighbour nodes, in this paper, *MT* function is proposed to be taken as the first objective function of I-NSGAII and its definition is introduced in the following.

**Definition 5 (*MT*).** Suppose $S(S_1,S_2,\ldots,S_u)$ is a community structure and $u$ is the number of communities for $S$, *MT* is defined as

$$MT=\sum_{j=1}^{u}\sum_{i=1}^{t}\frac{k_i^{S_j}}{k_i} \tag{5}$$

where $k_i^{S_j}$ is the internal degree of node $i$ belonging to community $S_j$, $k_i$ is the degree of node $i$, and $t$ is the node number of $S_j$.

Definition 5 indicates that *MT* function analyses community structure from the perspective of nodes within community; and the denser the intra-community connections, the higher the probability that each node shares the same community with its neighbour nodes.

The second objective function adopted in this paper is *AC* [16] function evaluating the sparsity of inter-community connections and its definition is provided in the following.

**Definition 6 (*AC*).** Suppose $S$ is a community structure, then conductance $\phi(S)$ of $S$ is defined as

$$\phi(S)=cs\,/\,min(Vol(S),Vol(V-S)) \tag{6}$$

where $cs=|\{a,\ b{:}a\in S,\ b\notin S\}|$ is number of edges connecting the node of community $S$ with external nodes, and $Vol(s)=\sum_{i\in s}k_i$ , $k_i$ is the degree of node $i$, then *AC* is defined as

$$AC=\frac{1}{u}\sum_{j=1}^{u}\phi(S_j) \tag{7}$$

where $u$ is the number of communities in a network and $S_j$ refers to the $j^{\mathrm{th}}$ community in a network.

Definition 6 indicates that AC function analyses community structure from the perspective of inter-community connections; and the sparser the inter-community connections, the lower the value of AC function.

These two objective functions reflect the advantages and disadvantages from different angles of community structure in complex networks. And that is consistent with our intuitive understanding of community structure—intra-community connections are dense and inter-community connections are sparse. The target of I-NSGAII is to obtain the community structure which is of relatively high *MT* value and low *AC* value.

### 3.2 Genetic representation and population initialization

#### 3.2.1 Genetic representation

The genetic representation is the strategy of locus-based adjacency representation [7] in I-NSGAII. In this representation, each individual $g$ in the population consists of $n$ genes $\{g_1, g_2, \ldots, g_n\}$. The allele value $j$ is the numerical value within the range $\{1\ldots n\}$. In the complex network G=(V, E), if the value of the $i^{\mathrm{th}}$ allele value is $j$, $i$ is linked with $j$. At the same time, it is meant that $i$ and $j$ belong to the same community. In the decoding step, all nodes with link are partitioned into the same community. The main advantage of this representation is that the number of communities included by each individual is automatically determined in decoding step. Fig.1 illustrates the strategy of locus-based adjacency representation for a network of 8 nodes.

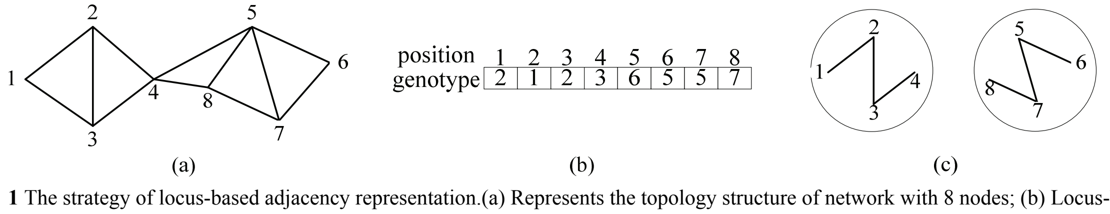


**Figure 1** The strategy of locus-based adjacency representation.(a) Represents the topology structure of network with 8 nodes; (b) Locus-based representation of an individual; (c) Graph-based structure of the individual

#### 3.2.2 Population initialization

The effective connections of the nodes in network are used as the initial population process in I-NSGAII algorithm, which means that the allele value of gene $i$ is its neighbours.

For example, the neighbour nodes of node 1 in Fig. 1(a) are node 2 and node 3. During the process of encoding, the corresponding allele value of gene 1 is chosen randomly from 2 and 3. And Fig. 1(b) illustrates the phenomenon of random selection of 2.

#### 3.2.3 Unified label for individual

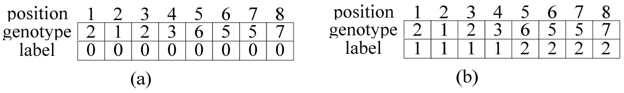
The above elaboration shows that the strategy of locus-based adjacency representation adopts gene $i$ and allele value $j$ to represent that link exists between node $i$ and node $j$. And during the iterative process of algorithm, it is of necessity to search the corresponding relationship between gene and allele value, decode each individual and calculate the objective function value of each individual. Thus, the running efficiency of algorithm is greatly reduced. In view of this, the strategy of unified label (U-L) is proposed in this paper to solve the problem.

The main idea of this strategy is that a label is attached to each gene and the same label value is set up to the corresponding labels of all the nodes (genes) belonging to the same community. The algorithm of U-L is described in the following.

```
Algorithm 1 process of U-L algorithm
Input  $R$ //individual to be unified labeled
Output  $R_{label}$ // individual unified labeled
Begin
  $label$←0 // initialize label value of community to 0
  $labelCom$←0
  for each $i \in n$ // $n$ is the gene numbers of individual $R$
    if $label(i)$==0 // $label(i)$ represents the community label value of gene $i$
      $findGene$←$i$ ; $g_i$ ←take out the allele value of gene $i$
      $gLa$←take out the community label value of gene $g_i$
      while $gLa$ ==0 && $g_i \notin findGene$
        $findGene$←$findGene \cup g_i$
        $g_i$←take out the allele value of gene $g_i$
        $gLa$←$label(g_i)$
      end while
      if $gLa$ >0 then set the community label of gene in set $findGene$ to be $gLa$
      else
        $labelCom$←$labelCom$+1
        set the community label value of gene in set $findGene$ to be $labelcom$
  end if ; end if ; end for
End
```

Fig. 2(a) attach the corresponding community label of gene on the basis that gene and allele value are reserved as in Fig. 1(b). Fig. 2(b) represents the state of individual after unified label; it indicates that the corresponding community of node could be determined directly by label.

From the above illustrations, it is noticeable that the corresponding community of gene after being unified labelled could be detected through its label value. And the decoding of each individual could be conducted directly without the need of repeatedly searching the corresponding relationship between gene and allele value. Thus, it is easier to calculate the objective function value.



| position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| genotype | 2 | 1 | 2 | 3 | 6 | 5 | 5 | 7 |
| label | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

(a)

| position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| genotype | 2 | 1 | 2 | 3 | 6 | 5 | 5 | 7 |
| label | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |

(b)

**Figure 2** Instance of unified label for individual. (a) Represents individual after being labelled; (b) Represents individual after unified labelling.

## 3.3 Genetic operator

### 3.3.1 One-way crossover strategy

The traditional crossover operators include one-point crossover and two-point crossover etc. And those operators are of vigorous randomicity, thus the present good community structure could be easily destroyed. In view of this, one-way crossover strategy [9] is taken as the crossover strategy of I-NSGAII in this paper.

**Definition 7 (One-way crossover)**. Suppose $a$, $b$ are two individuals, $a$ is source individual and $b$ is destination individual. One-way crossover operation is defined as follows: first, one certain node (gene) $i$ and its corresponding community label value $La(i)$ are randomly selected from individual $a$, then the label value equivalent to $La(i)$ of a is assigned to the corresponding gene's community label of destination individual $b$. It is formulated as

$$La(i) \to Lb(x), \forall x \in \{x \mid La(x) = La(i)\} \tag{8}$$

Tab. 1 provides an instance of one-way crossover selecting node 4. It indicates that one-way crossover operation is able to reserve good community structure, and meanwhile the generated child individuals carry the characteristics of parent individual.

**Table 1** Instance of one-way crossover

| v | | a(Source) | b(Destination) | new |
|---|---|---|---|---|
| 1 | | (1) → | 1→ | (1) |
| 2 | | 2 | 1 | 1 |
| 3 | | 2 | 2 | 2 |
| (4) | → | (1) → | 3→ | (1) |
| 5 | | 3 | 3 | 3 |
| 6 | | 3 | 3 | 3 |
| 7 | | (1) → | 1→ | (1) |
| 8 | | 3 | 2 | 2 |

### 3.3.2 Local mutation strategy

Local mutation strategy is adopted in this paper. Its main idea is that individual mutation is achieved by compelling mutated node to be in the same community with its most neighbour nodes. And local mutation strategy is defined as follows.

**Definition 8 (local mutation)** [10]. Suppose $a$ is an individual, $La(i)$ is community label value of gene $i$ in individual $a$, then local mutation is defined as that community label value $La(i)$ of gene $i$ is mutated into community label value of most neighbour nodes in the community, i.e.

$$La(i) \leftarrow \arg\max_{l} \sum_{u \in n(i)} \delta(La(u), l) \tag{9}$$

where $La(u)$ is the present community label value of node $u$, $n(i)$ is the set of node $i$'s neighbor nodes.

### 3.3.3 Diversity evolutionary strategy

For the purpose of enhancing the diversity of the set of Pareto optimal solutions of multi-objective community detection, and avoid the local optimization of the set of Pareto optimal solutions, diversity evolutionary strategy (DES) suitable to the field of multi-objective community detection is proposed in this paper.

**Definition 9 (overall mutation for neighbor nodes)**. Suppose $a$ is an individual, $La(i)$ is the corresponding community label value of gene $i$ in individual $a$, then value of $La(i)$ could be randomly changed into the community label value of node $i$'s neighbour nodes, i.e.

$$La(i) \leftarrow La(j), \forall j \in n(i) \tag{10}$$

where $n(i)$ is the set of node $i$'s neighbour nodes, corresponding label values of all genes in individual $a$ are mutated universally and finally a new individual $M_a$ is obtained, then the process is called overall mutation for neighbour nodes.

To sum up, traditional mutation only changes a small number of genes in individual. Different from that, overall mutation for neighbour nodes completely changes individuals according to the topology structure information of network. And no maternal information is maintained in changed individual. Thus, the searching space of solution could be effectively expanded.

Based on overall mutation for neighbour nodes, diversity evolutionary strategy is proposed in this paper. In population $P$, $e$ individuals are randomly selected from every front of the set of Pareto optimal solutions and they compose set $T$ whose length is $E_m$. Then individual $a$ is selected in turn from $T$ and the operation of overall mutation for neighbour nodes is conducted on it. Finally, $E_m$ new individuals are obtained.

The purpose of this strategy is, on the basis that a certain number of individuals of the set of Pareto optimal solutions are kept stable, take out some individuals and conduct on them the operation of overall mutation for neighbour nodes that is able to enlarge the searching space of solutions. Thus, the diversity of the set of Pareto optimal solutions is enhanced and the searching space of Pareto optimal solutions is enlarged while the present set of Pareto optimal solutions is kept stable. Consequently, local optimization of the Pareto optimal solutions is avoided. The illustration of DES is presented as follows.

Algorithm 2 process of DES

```
Input  $h$, $P_{size}$ , $E_m$ //$h$ is the number of Pareto front, $P_{size}$ is the size of population, and $E_m$ is the maximum size of overall mutation for neighbor nodes
Output  $P_{des}$ // the population generated after the operation of DES algorithm
Begin
W←$E_m$ / $h$
for $i$=1:$h$
    $TP$←$W$ individuals are randomly selected from the set of Pareto optimal solutions of the $i$th Pareto front
    $EP \leftarrow EP \bigcup TP$
end for
for each $R \in EP$
    for $j$=1:$n$  //$n$ is the gene number of individual $R$
        $V_j$←the set of node $j$'s neighbor nodes
        $La(j)$←the community label value of any node in $V_j$
end for ; end for
End
```

### 3.4 Description of I-NSGAII algorithm

Under the framework of NSGAII and based on the strategy of locus-based adjacency representation, I-NSGAII adopts the method of effective connections of the nodes in network to initialize population, and applies the strategy of unified label to label each gene of individual in population. Then, it adopts the strategies of one-way crossover, local mutation, and diversity evolutionary to optimize the individuals in population; moreover, the strategies of Pareto dominated sorting and crowding distance assignment are utilized to select excellent solutions; finally, the set of Pareto optimal solutions is obtained. The description of I-NSGAII algorithm is presented as follows.

Algorithm 3 process of I-NSGAII

```
Input  $N$, $P_{size}$ //$N$ is complex networks, $P_{size}$ is the size of population
Output  $P_{i\text{-}nsgaii}$ // the finally generated set of Pareto optimal solutions regarding the community structure of network
Begin
$P_0$←initialize the population and operation of unified label
$P_0^{new}$←conduct on population the operation of one-way crossover and local mutation
t=0
main loop
    $R_t \leftarrow P_t \cup P_t^{new}$
    $MT(R_t)$ and $AC(R_t)$ //calculate the $MT$ and $AC$ value of each individual
    $R_t$←conduct the operation of Pareto dominated sorting and crowding distance assignment
    $t$=$t$+1
    $P_t$←the first $P_{size}$ individuals from $R_t$
    $P_t$←conduct on population the operation of one-way crossover and local mutation
    $P_t^{new}$←conduct on population the operation of diversity evolutionary
end loop
End
```

### 3.5 Analysis of time complexity

Suppose the node number in network is $n$, node's average degree is $k$, population size is $L$, and the number of DES algorithm's overall mutation for neighbour nodes is $E_m$. The analysing process is as follows.

In I-NSGAII algorithm, the strategies of population initialization, unified label, one-way crossover and local mutation are adopted to conduct operations on every gene of each individual in population. The time complexity is smaller than $O(kn)$ when it conducts operation on single individual, and the time complexity is smaller than $O(Lkn)$ when it conducts operation on all individuals of population. The time complexity is $O(kn)$ when the operation of overall mutation for neighbour nodes is conducted on single individual and the running time is $O(E_m\ kn)$ when the operation is conducted on $E_m$ individuals. Research shows [8] that the time complexity of crowding distance assignment is $O(M(2n)\log(2n))$ and that of Pareto dominated sorting is $O(Mn^2)$, where $M$ is the number of objective functions. The above analysis indicates that only the time complexity of Pareto dominated sorting is $O(Mn^2)$ and the time complexities of others are smaller than $O(n^2)$. Because $M$ is constant, the time complexity of I-NSGAII algorithm, proposed in this paper, is $O(n^2)$.

## 4 Experiment results

The algorithm of I-NSGAII proposed in this paper is tested on synthetic and real-world networks. Its validity and feasibility are verified through the comparisons with classic algorithms of single-objective community detection such as FN [4], LPA [17], and GANET [13], and with algorithms of multi-objective community detection such as NSGAII、MOGANET [7]. What needs noting is that all the operators of I-NSGAII are identical with those of NSGAII except that NSGAII does not apply diversity evolutionary strategy. Tab. 2 presents the parameter setting of I-NSGAII algorithm.

**Table 2** Parameter setting of I-NSGAII algorithm

| Parameter | Meaning | Value |
|---|---|---|
| $P_{size}$ | Population size | 100 |
| $L$ | Iterative time | 50 |
| $P_c$ | Crossover rate | 0.6 |
| $P_m$ | Mutation rate | 0.4 |
| $E_m$ | Size of overall mutation for neighbour nodes | 10 |

Two classic evaluation criteria are adopted in this paper to evaluate the performances of various algorithms and measure the advantages and disadvantages of these algorithms. The first evaluation criterion is normalized mutual information (NMI) [18] evaluating the accuracy rate of community detection. The second evaluation criterion is modularity ($Q$) evaluating the density of intra-community connections.

### 4.1 Experiment results of synthetic networks

LFR benchmark [19] proposed by Lancichinetti is adopted here as synthetic networks. LFR benchmark has a statistical property that most real-world networks seem to share. Thus, the LFR benchmark is adopted as the experimental network of I-NSGAII algorithm.

The parameters of LFR benchmark are set as follows. Network size is $N = 100 \div 500$, average degree of node is $k = 15$, maximum degree of node is $k_{\max} = 50$, mixing parameter $\gamma$ (each node shares a fraction $\gamma$ of its edges with nodes in other communities) indicates that the higher the $\gamma$ value, the vaguer the community structure in network, and it is harder for algorithm to detect community structure. When $\gamma > 0,5$, there is no community structure in network, thus the range of parameter $\gamma$ is set to be $0,1 \div 0,5$.

#### 4.1.1 Diversity analysis

The diversity of the set of Pareto optimal solutions is analysed in this section to test whether the diversity evolutionary strategy is effective. Two network structures are adopted here whose mixing parameter values are respectively $\gamma = 0,2$ and $\gamma = 0,4$, and these two network structures share the same network size $N = 200$. According to the property of LFR benchmark, $\gamma = 0,2$ indicates that the community structure is clearer and $\gamma = 0,4$ indicates that it is vaguer.

When $\gamma = 0,2$, the distributions of corresponding Pareto front of the set of Pareto optimal solutions produced by I-NSGAII and NSGAII algorithms are presented in Fig. 3(a) which indicates that the distribution produced by I-NSGAII algorithm is broader. And when $\gamma = 0,4$, the distributions of corresponding Pareto front of the set of Pareto optimal solutions produced by I-NSGAII and NSGAII algorithms are presented in Fig. 3(b) indicating that the set of Pareto optimal solutions of NSGAII algorithm is surrounded by that of I-NSGAII algorithm. To sum up, the distribution of the set of Pareto optimal solutions obtained by I-NSGAII algorithm is broader no matter whether the community structure is clearer or vaguer. Thus, it is verified that the diversity of the set of Pareto optimal solutions obtained by I-NSGAII algorithm is stronger. In addition, it is testified that diversity evolutionary strategy is able to enhance the diversity of the obtained set of Pareto optimal solutions.

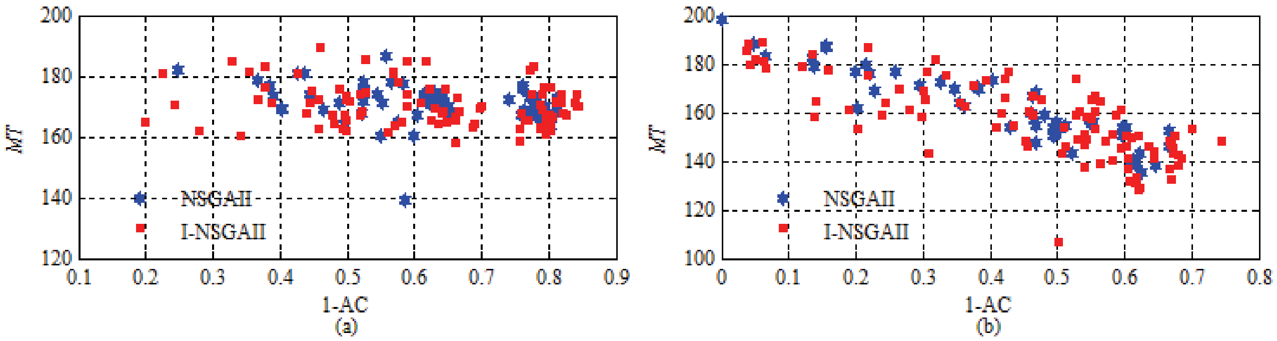


**Figure 3** Comparison of the diversity of the set of Pareto optimal solutions between I-NSGAII and NSGAII. (a) Distribution of two algorithms' Pareto front when $\gamma = 0,2$ ;(b) Distribution of two algorithms' Pareto front when $\gamma = 0,4$.

#### 4.1.2 Comprehensive analyses of I-NSGAII

Under the conditions that network size is $N = 200$, algorithms of I-NSGAII, NSGAII and MOGANET are to be run 10 times for each. All the optimal average values of *NMI* and $Q$ obtained by I-NSGAII, NSGAII and MOGANET are compared with the optimal values of *NMI* and $Q$ obtained by FN, LPA and GANET. And the comparative results are presented respectively in Fig. 4(a) and Fig. 4(b).

Fig. 4(a) indicates that the obtained *NMI* values of all those algorithms are very close when $\gamma = 0,1$, and the obtained *NMI* values of those algorithms except MOGANET are still high when $\gamma$ = 0,2. Under the conditions that $\gamma$ = 0,3 ÷ 0,5, community structure becomes vague gradually and detection difficulty becomes increased, the *NMI* values of the other algorithms decline dramatically as compared to the *NMI* value of I-NSGAII. Thus, compared with the other algorithms, I-NSGAII is able to detect community structure more accurately.

Fig. 4(b) indicates that the obtained *Q* values of all those algorithms are very close when $\gamma$ = 0,1, and the obtained *Q* value of I-NSGAII is a little bit lower than that of FN algorithm which is based on the optimization of modularity when $\gamma$ = 0,2 . In addition, when $\gamma$ = 0,3 ÷ 0,5, the obtained *Q* value of I-NSGAII is much higher than that of the others. Thus, I-NSGAII is able to detect denser community structure.

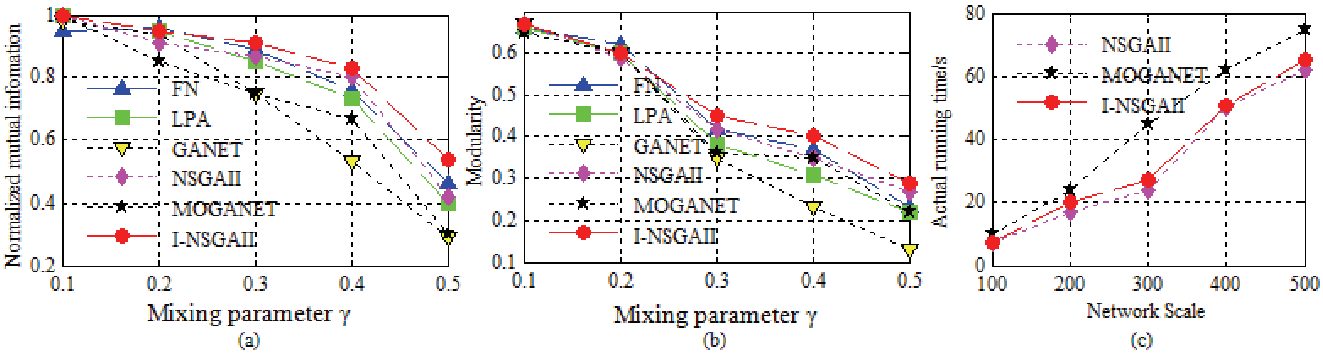


**Figure 4** Compare I-NSGAII with other algorithms on the LFR benchmark. (a) The comparative result of algorithms' NMI values; (b) The comparative result of algorithms' *Q* values; (c) Comparative result of running time between I-NSGAII, MOGANET and NSGAII

### 4.1.3 Analyses of time efficiency

Time efficiency of algorithm is another important evaluation criterion of community detection in complex networks. Here, for a further evaluation of I-NSGAII's performance, I-NSGAII algorithm is compared and analysed with algorithms of NSGAII and MOGANET. Note that no comparison is conducted between I-NSGAII algorithm and other single-objective community detection algorithms, which is mainly because that algorithm of multi-objective community detection simultaneously optimizes multiple objective functions, and that optimization is usually time-consuming. Thus, community detection algorithm based on single-objective optimization is usually superior to the algorithm of multi-objective community detection in terms of time efficiency.

Due to the different abilities of the evolutionary strategies adopted by each multi-objective community detection algorithm, the iterative times selected by each algorithm are also different. For a fair analysis of algorithm's time efficiency, that each algorithm's *NMI* is higher than 80 % is regarded as the terminal condition of iteration. Fig. 4(c) presents the changing trend of running time along with the change of network size.

The Figure shows that the running time of I-NSGAII algorithm is almost equal to that of NSGAII algorithm while superior to that of MOGANET algorithm. As the I-NSGAII algorithm strengthened the diversity of solutions by adding the diversity evolutionary strategy, it takes a bit more time for the optimal searching. Nevertheless, the result indicates that the time-consuming of I-NSGAII algorithm with diversity evolutionary strategy is rather small.

## 4.2 Experiment results of real-world networks

Due to the differences in topology property of synthetic and real-world networks, four classical real-world networks [20] whose community structures are already known are adopted in this paper to perform a further test on I-NSGAII algorithm. A description of them is given in Tab. 3.

**Table 3** Real-world networks

| Networks | Nodes | Edges | Description |
|---|---|---|---|
| Karate | 34 | 78 | Zachary's karate club |
| Dolphins | 62 | 159 | Dolphin social network |
| Polbooks | 105 | 441 | Books about US politics |
| Football | 115 | 613 | American College football |

Tab. 4 shows the comparative results of *NMI* values obtained of those algorithms, where MAXNMI represents optimal *NMI* values by running algorithms 10 times and AVGNMI represents the average value of the 10-times optimal *NMI* values. It is indicated that the *NMI* values of community structures detected by running I-NSGAII algorithm on these four real-world networks are obviously better than the values obtained by the other 5 algorithms.

**Table 4** *NMI* value comparisons between I-NSGAII and the other 5 algorithms

| *NMI* | FN | LPA | GANET | MOGANET | | NSGAII | | I-NSGAII | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *MAXNMI* | *AVGNMI* | *MAXNMI* | *AVGNMI* | *MAXNMI* | *AVGNMI* |
| Karate | 0,6925 | 0,8372 | 0,5998 | 1 | 0,9186 | 1 | 0,9186 | 1 | 0,9853 |
| Dolphins | 0, 5727 | 0,5914 | 0,6981 | 0,9065 | 0,8704 | 1 | 0,9092 | 1 | 0,9332 |
| Polbooks | 0,5308 | 0,5307 | 0,5930 | 0,6020 | 0,5741 | 0,6046 | 0,5834 | 0,6809 | 0,6397 |
| Football | 0,7571 | 0,8398 | 0,6656 | 0,7554 | 0,7202 | 0,7448 | 0,7021 | 0,8281 | 0,7874 |

The distributions of the corresponding Pareto front of the set of Pareto optimal solutions obtained by running I-NSGAII algorithm on Karate network are presented in Fig. 5(a) in which the *NMI* values of the corresponding community structures of some nodes are listed. Fig. 5(b) shows the community structure when *NMI* value is 1, which indicates that the community structure detected by I-NSGAII algorithm is identical to true community structure. Fig. 5(c) shows the community structure when *NMI* value is 0,7172, which indicates that the community on the right is divided into two small communities by I-NSGAII algorithm. Fig. 5(d) shows the community structure when *NMI* value is 0,6021, which indicates that two relatively large communities are divided into four relatively small communities by I-NSGAII algorithm. To sum up, the set of Pareto optimal solutions obtained by I-NSGAII algorithm is of better hierarchy and the algorithm allows us to analyse community structure within different hierarchies.

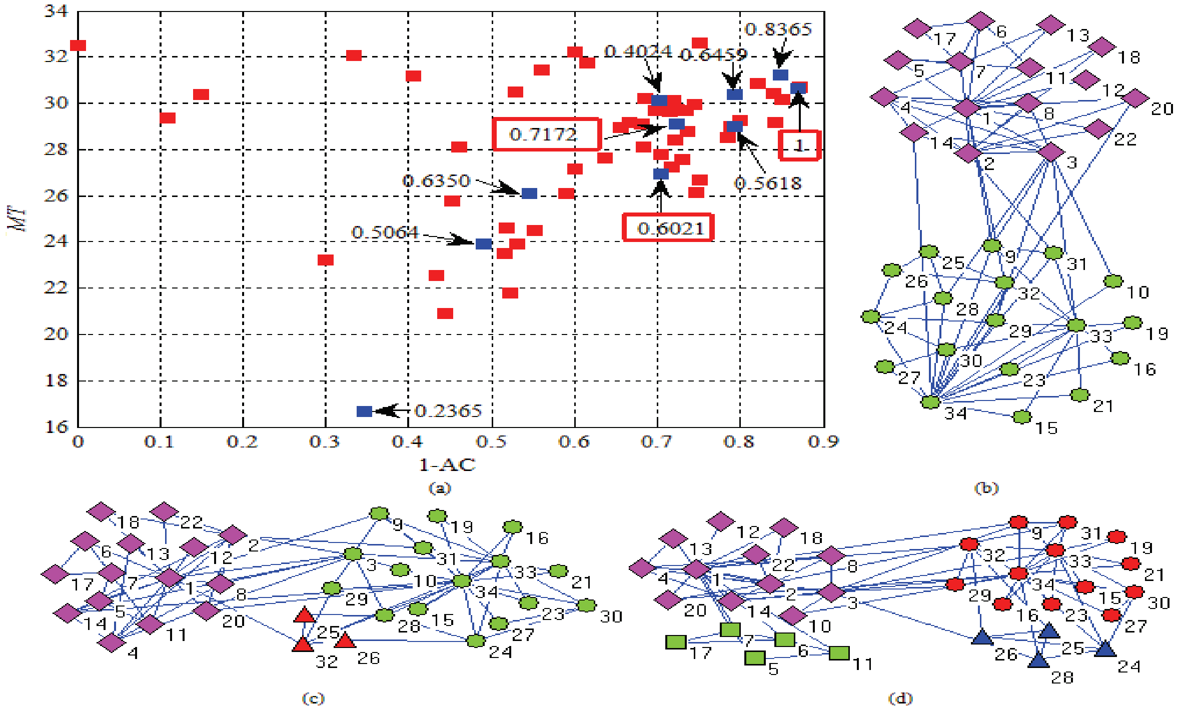


**Figure 5** I-NSGAII's result of community detection on Karate. (a) Pareto front produced by running I-NSGAII; (b) Community structure corresponding to *NMI* = 1; (c) Community structure corresponding to *NMI* = 0,7172; (d) Community structure corresponding to *NMI* = 0,6021.

Tab. 5 shows the comparative results of *Q* values obtained by each algorithm on these four real-world networks, where MAXQ represents the 10-times optimal *Q* values and AVGQ represents the average value of the 10-times optimal *Q* values.

It is indicated that on Karate network the maximum *Q* values obtained by running I-NSGAII algorithm are equal to the *Q* value obtained by running MOGANET; all the Q values obtained by running I-NSGAII algorithm on Dolphins, Polbooks and Football networks are higher than the *Q* values obtained by the other 5 algorithms.

**Table 5** *Q* value comparisons between I-NSGAII and the other 5 algorithms

| *Q* | FN | LPA | GANET | MOGANET | | NSGAII | | I-NSGAII | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *MAXQ* | *AVGQ* | *MAXQ* | *AVGQ* | *MAXQ* | *AVGQ* |
| Karate | 0,3807 | 0,4020 | 0,3998 | 0,4198 | 0,4050 | 0,4156 | 0,3993 | 0,4198 | 0,4056 |
| Dolphins | 0,4897 | 0,5113 | 0,4824 | 0,5050 | 0,4857 | 0,4876 | 0,4811 | 0,5265 | 0,5182 |
| Polbooks | 0,5020 | 0,5156 | 0,4906 | 0,5209 | 0,5169 | 0,4852 | 0,4692 | 0,5264 | 0,5226 |
| Football | 0,5773 | 0,5725 | 0,5840 | 0,5058 | 0,4938 | 0,5637 | 0,5413 | 0,5888 | 0,5531 |

## 5 Conclusion

An effective method of multi-objective community detection, named I-NSGAII is proposed in this paper. This method, based on intuitive understanding of community, first simultaneously optimizes two objective functions of *MT* and *AC* which respectively evaluates the density of intra-community connections and the sparsity of inter-community connection. These two objective functions, from different perspectives, analyse the community structure in complex networks, and obtains a set of Pareto optima solutions with each of these solutions correspond to a trade-off between these two objective functions. Then, I-NSGAII algorithm, based on the framework of NSGAII, adopts diversity evolutionary strategy which evenly selects several individuals from every Pareto front of the set of Pareto optimal solutions and changes all the genes of individuals under the guidance of network structure. And that enables the algorithm to search the set of solutions from broader solution space, thus enhances the diversity of the set of Pareto optimal solutions and avoids local optimization of the set of Pareto optimal solutions. Last, to improve algorithm's searching ability of optimal solutions, this paper proposes the integrated application of genetic strategies which are more suitable to the field of community detection such as locus-based adjacency representation, unified label, one-way crossover and local mutation.

I-NSGAII algorithm is tested both on synthetic and real-world networks and compared with multiple classic algorithms. The experiment results show that the set of Pareto optimal solutions is of strong diversity and able to demonstrate community structure in complex networks of diverse hierarchies. In addition, evaluations of community structure obtained by I-NSGAII algorithm are conducted from various perspectives and those evaluations illustrate the validity of I-NSGAII algorithm. Summarization of the above results shows that I-NSGAII algorithm is of certain advantages on community structure detection in complex networks.

### Acknowledgments

This work is sponsored by the National Natural Science Foundation of China (61370083, 61073043, 61073041), the National Research Foundation for the Doctoral Program of Higher Education of China (No. 20112304110011, No. 20122304110012), the Natural Science Foundation of Heilongjiang Province (F200901), and the Harbin Outstanding Academic Leader Foundation of Heilongjiang Province of China (No. 2011RFXXG015).

## 6 References

[1] Watts, D. J.; Strogatz, S. H. Collective dynamics of 'small-world' networks. // Nature. 393, 84(1998), pp. 440- 442. DOI: 10.1038/30918

[2] Adamic, L. A.; Huberman, B. A.; Barabasi, A. L.; Albert. R.; Jeong. H.; Bianconi. G. Power-law distribution of the World Wide Web. // Science. 287, 5461(2000), pp. 2115a. DOI: 10.1126/science.287.5461.2115a

[3] Newman, M. E. J.; Girvan, M. Finding and evaluating community structure in networks. // Physical Review E, 69, 2(2004), pp. 026113. DOI: 10.1103/PhysRevE.69.026113

[4] Newman, M. E. J. Fast algorithm for detecting community structure in networks. // Physical Review E. 69, 6(2004), pp. 066133. DOI: 10.1103/PhysRevE.69.066133

[5] Blondelv, D.; Guillaume, J. L.; Lambiottee, R.; Lefebvre,E. Fast unfolding of communities in large networks. // Journal of Statistical Mechanics: Theory and Experiment. 10, (2008), pp. P10008. DOI: 10.1088/1742-5468/2008/10/P10008

[6] Shi, C.; Yan, Z. Y.; Cai, Y. N.; Wu, B. Multi-objective community detection in complex networks. // Applied Soft Computing. 12, 2(2012), pp. 850-859. DOI: 10.1016/j.asoc.2011.10.005

[7] Pizzuti, C. A Multiobjective genetic algorithm to find communities in complex networks. // IEEE Transactions on Evolutionary Computation. 16, 3(2012), pp. 418-430. DOI: 10.1109/TEVC.2011.2161090

[8] Deb, K.; Pratap, A.; Agarwal, S. Meyarivan T. A. Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. // IEEE Transactions on Evolutionary Computation. 6, 2(2002), pp. 182-197. DOI: 10.1109/4235.996017

[9] Tasgin, M.; Herdagdelen, A.; Bingol, H. Community detection in complex networks using genetic algorithms. // arXiv preprint arXiv, 0711.0491, (2007), URL: http://arxiv.org/abs/0711.0491v1, 2006.

[10] He, D. X.; Zhou, X.; Wang, Z.; Zhou, C. G.; Wang, Z.; Jin, D. Community mining in complex networks-clustering combination based genetic algorithm. // Acta Automatica Sinica. 36, 8(2010), pp. 1160-1170. DOI: 10.3724/SP.J.1004.2010.01160

[11] Girvan, M.; Newman, M. E. J. Community structure in social and biological networks. // Proceedings of the National Academy of Science. 99, 12(2002), pp. 7821-7826. DOI: 10.1073/pnas.122653799

[12] Hu, Y. Q.; Li, M. H.; Zhang, P. Community detection by signalling on complex networks. // Physical Review E. 78, 1(2008), pp. 016115. DOI: 10.1103/PhysRevE.78.016115

[13] C. Pizzuti, GA-NET: a genetic algorithm for community detection in social networks. // Proceedings of the 10th International Conference on Parallel Problem Solving from Nature / Dortmund, 2008, pp. 1081-1090. DOI: 10.1007/978-3-540-87700-4_107

[14] Gong, M. G.; Ma, L. J.; Zhang, Q. F.; Jiao, L. C. Community detection in networks by using multiobjective evolutionary algorithm with decomposition. // Physica A: Statistical Mechanics and its Applications. 391, 15(2012), pp. 4050-4060. DOI: 10.1016/j.physa.2012.03.021

[15] Huang, F. L.; Zhang, S. C.; Zhu, X. F. Discovering network community based on multi-objective optimization. // Journal of Software (in Chinese), (2013), URL:http://www.jos.org.cn/1000-9825/4400.htm

[16] Leskovec, J.; Lang, K. J.; Mahoney, M. W. Empirical comparison of algorithms for network community detection. // Proceedings of the 19th International World Wide Web Conference (WWW'10)/New York, 2010, pp. 631-640. DOI: 10.1145/1772690.1772755

[17] Raghavan, U. N.; Albert, R.; Kumara, S. Near linear-time algorithm to detect community structures in large-scale networks. // Physical Review E. 76, 3(2007), pp. 036106. DOI: 10.1103/PhysRevE.76.036106

[18] Danon, L.; Diaz-Guilera, A.; Duch, J.; Arenas, A. Comparing community structure identification. // Journal of Statistical Mechanics: Theory and Experiment. 9 (2005), pp. P09008. DOI: 10.1088/1742-5468/2005/09/P09008

[19] Lancichinetti, A.; Fortunato, S.; Radicchi, F. Benchmark graphs for testing community detection algorithms. // Physical Review E.78, 4(2008), pp. 046110. DOI: 10.1103/PhysRevE.78.046110

[20] Newman, M. E. J. Network data.2006.URL:http://www-personal.umich.edu/~mejn/netdata/ (02.25.2014)

### Authors' addresses

*Kun Deng*
College of Computer Science and Technology,
Harbin Engineering University,
Harbin 150001, China
E-mail: dengkun@hrbeu.edu.cn

*Jian Pei Zhang*
College of Computer Science and Technology,
Harbin Engineering University,
Harbin 150001, China
E-mail: zhangjianpei@hrbeu.edu.cn

*Jing Yang*
College of Computer Science and Technology,
Harbin Engineering University,
Harbin 150001, China
E-mail: yangjing@hrbeu.edu.cn