International Journal of Grid and Distributed Computing
Vol. 9, No. 10 (2016), pp.357-374
http://dx.doi.org/10.14257/ijgdc.2016.9.10.32

Community Detection in Complex Networks based on Improved Genetic Algorithm and Local Optimization

Kun Deng, XingYan Liu and WenPing Li

College of Mathematics Physics and Information Engineering, JiaXing University, JiaXing 314001, China
dengkun@hrbeu.edu.cn

Abstract

This paper proposes the community detection in complex networks based on improved genetic algorithm and local optimization (IGALO) in terms of the defect that traditional community detection approaches based on genetic algorithm have strong randomness and weak searching ability in the process of community detection. Taking modularity function Q as the objective function, IGALO algorithm adopts label propagation method of one-iteration to initialize population so as to generate initial population with certain precision. Then, anti-destructive one-way crossover strategy is proposed to ensure the crossover operation to develop in the direction of making community structure increase to modularity function. Finally, mutation strategy of node local optimization is proposed to improve the searching efficiency of algorithm. This algorithm effectively overcomes the defect that traditional algorithms have weak searching ability and improves the community detection accuracy. Tests are made on benchmark networks and real-world networks and comparative analysis is also made with various classic algorithms. The results show that IGALO algorithm is effective and feasible.

Keywords*: complex networks, community detection, genetic algorithm, anti-destructive one-way crossover strategy, mutation strategy of node local optimization*

1. Introduction

Currently, the study of complex networks has been the hotspot in research field and it also draws much attention of researchers from the fields of computer, mathematics, physics, and sociology and so on. Through continuous study of physical meaning and mathematical characteristics of network features, the property of community structure is found in complex networks, which refers to the density of intra-community connections and the sparsity of inter-community connections [1]. Community detection of complex networks aims to reveal the community structure existing in networks, which has both theoretical significance and practical significance for the analysis of topological structure of complex networks, the discovery of implicit laws in networks and the analysis of network evolution.

In recent years, various classic approaches to community detection have been proposed by scholars from different fields, such as GN algorithm based on division [2], CPM algorithm based on clique percolation [3], LC algorithm based on link clustering [4], LPA algorithm based on label propagation [5], LFM algorithm based on seed expansion [6], BGLL algorithm based on modularity optimization [7], *etc.*, At present, the approach to community detection based on modularity function *Q* has become one of the main stream algorithms in community detection. Actually, this method transfers the problem of community detection to the problem of function optimization. The study shows [8] that to calculate the maximum modularity is NP-complete problem. Thus, as the most effective method of solving NP problem, genetic algorithm (GA) has been widely used in the field of community detection. For example, GANET algorithm [9], proposed by Pizzuti *et. al.*,

ISSN: 2005-4262 IJGDC

introduces locus-based adjacency representation into community detection and adopts the strategies of uniform crossover, random mutation and elite selection to complete the task of community detection. CDGA algorithm [10] is proposed by Tasgin *et al.* And in order to maximally reserve the optimal community structure of individuals, it puts up with the strategy of one-way crossover and completes community detection with the combination of strategies of neighbor nodes initial population and random mutation. On the basis of CDGA algorithm, Shi *et. al.*, [11] introduce the one-way crossover operation based on the strategy of string encoding into the encoding strategy based on graphs and designs genetic algorithm which can deal with large-scale networks. CCGA algorithm [12], proposed by He *et. al.*, views string encoding method as the encoding method of algorithm and adopts the method of Markov random walk to initialize the population. And it also proposes the strategy of multi-individual clustering combination crossover with the combination of function Q and completes the task of community detection with the combination of local mutation strategy. Meme-Net algorithm [13] is proposed by Gong *et. al.*, In order to obtain better community structure, hill climbing algorithm is used to further process the elite individuals obtained through genetic algorithm. GALS algorithm [14], proposed by Liu *et. al.*, proposes local search mutation strategy on the basis of proved local monotonicity of function Q and completes the task of community detection in complex networks in combination with the locus-based adjacency encoding strategy and uniform crossover strategy. MIGA algorithm [15], proposed by Shang *et. al.*, optimizes the elite individuals through simulated annealing algorithm on the basis of genetic algorithm, so as to obtain better community structure. Although approaches to community detection based on genetic algorithm are widely applied, the defects of slow convergence speed and easily falling into local optimal solution still exist, so the quality of community detection will be influenced. It is analyzed that the reason lies in the randomness of algorithm operation and the destructive strategies of crossover and mutation.

To solve the above problems, it is very necessary to explore the method of initial population with high precision and the reasonable crossover and mutation strategies. Thus, this paper proposes the community detection in complex networks based on improved genetic algorithm with local optimization (IGALO). The specific innovative points are as follows:

(1) Label propagation method of one-iteration is used to initialize the population, so as to improve the precision of individuals in initial population.

(2) Anti-destructive one-way crossover strategy is proposed in terms of the defect that traditional crossover strategy easily destroys the existing good community structure.

(3) This paper comes up with the mutation strategy of node local optimization, faced with the problem that traditional mutation strategy has strong randomness and weak local searching ability.

(4) IGALO algorithm is designed on the basis of the above strategies.

2. IGALO Algorithm

IGALO algorithm proposed in this paper views modularity function *Q* as the objective function and adopts string encoding method as encoding method of algorithm. IGALO Algorithm first generates initial population with certain precision through label propagation method of one-iteration. Then, anti-destructive one-way crossover strategy and mutation strategy of node local optimization are used for genetic evolution. Finally, the elite selection strategy of selected references [16] is adopted to choose the best individuals to complete the task of community detection.

2.1. Objective Functions

Modularity function *Q* is a quantitative index of evaluating the quality of community structure of network. Currently, this index has been generally accepted by researchers in

the field of community detection. Thus, function *Q* is the objective function of IGALO algorithm in this paper. Function *Q* aims to evaluate the difference between community structure obtained and random networks with stable node's degree. The higher value of function *Q* indicates a denser intra-community connection and a more qualified community structure. On the contrary, the quality of community structure is lower. Modularity function *Q* can be defined as:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{1}$$

Here e_{ii} indicates the proportion of the links with two nodes in the same community *i* to all links in network. And a_i indicates the proportion of links with at least one node in the same community *i* to all links in network.

2.2. Encoding Method

The two most commonly used encoding methods are string encoding method and locus-based adjacency encoding method, when genetic algorithm is used to complete the task of community detection [17]. Because of the simple and efficient features of string encoding method, this paper chooses string encoding method as the encoding method of IGALO algorithm. If *n* nodes exist in a certain complex network, each individual (chromosome) consists of *n* genes, which can be defined as follows:

$$Ind = \{l_1, l_2, ..., l_n\} \tag{2}$$

Here, l_i is the belonging community label of ith node. If $l_i = l_j$, then node *i* and node *j* belong to the same community. For example, there are 6 nodes ($v_1, v_2, v_3, v_4, v_5, v_6$) in complex networks and if they are represented as (1,2,2,1,2,2) in the method of string encoding, then it shows that node v_1 and node v_4 belong to the same community and that node v_2, node v_3, node v_5 and node v_6 belong to the same community. It can be seen that each individual represents a result of community detection when genetic algorithm is used to complete community detection.

2.3. Initial Population

Only when each individual of initial population has certain precision, the searching advantage of genetic algorithm can be fully realized. Thus, this paper adopts the idea based on label propagation [5] and generates initial population using the label propagation method of one-iteration. The idea is that in a complex network each node should belong to the same community as its most neighbor nodes, which can be defined as follows:

$$l_i = \arg\max_l \sum_{j \in adj(i)} \delta(l_j, l) \tag{3}$$

Here, $adj(i)$ is the set of neighbor nodes of node i.

Based on this, the label propagation process of one-iteration used in this paper is as follows:

(1) Initialize the label of all nodes in network.

(2) In the method of asynchronous updates, label of each node in network will have label update only once according to Formula (3).

It can be seen that when the label propagation method of one-iteration is used for community detection, labels connecting close nodes are easily to reach a consensus and nodes with the same label will belong to the same community. Thus, initial population with certain precise will be generated.

2.4. Crossover Strategy

When string encoding method is used to indicate community structure, the community that nodes in network belong to is just an identifier. And there is such a problem that multi-individuals may correspond to the same detection result. For example, there are 6 nodes ($v_1, v_2, v_3, v_4, v_5, v_6$) in a certain network. The community detection results of two individuals are respectively expressed as (1,2,3,1,2,3) and (4,5,6,4,5,6). And both detection results they indicate are node v_1 and node v_4 belong to the same community, node v_2 and node v_5 belong to the same community and that node v_3 and node v_6 belong to the same community. Obviously, each community label in each individual is randomly expressed. If traditional single-point crossover approach is adopted, the two individuals may become (1, 2, 3, 4, 5, 6) and (4, 5, 6, 1, 2, 3) after crossover. It shows that this crossover destroys the original existing community structure, which makes great impact on the searching ability of genetic algorithm. Thus, Tasgin proposed one-way crossover operation. The basic idea is to suppose *A* and *B* are two individuals before crossover, *A* is source individual and *B* is destination individual. L_i, the label of any gene i in individual *A*, is marked. Then assign all labels which are L_i in source individual *A* to labels of corresponding genes in destination individual *B*. One-way crossover strategy can solve the problems in traditional crossover strategy, but there is still the problem that good community structure is easily destroyed. As is illustrated in Figure 1, obviously node 1, 2, 3, 4, 5 belong to the same community and node 6, 7, 8 belong to the same community.

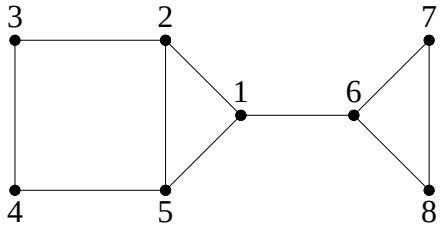

Figure 1. Example of Network with Two Communities

Combined with Figure 1, one-way crossover operation is showed in Table 1 when node 2 is chosen. It can be seen that all labels which are 1 in source individual A will have a one-time assignment to the corresponding genes in destination individual B. Thus, the new individual B^{new} will be generated. And nodes 2, 3, 4, 6, 7 will be detected as one community, which will destroy the better community structure that nodes 2, 3, 4, 5 in individual B belong to the same community. The reason for this is that one-way crossover has a one-time assignment for all the same labels in a certain community in source individual to destination individual, regardless of the effect on community structure in destination individual. As a result, destination individuals that have obtained better community structure usually will be destroyed. Focusing on this problem, this paper proposes anti-destructive one-way crossover strategy.

According to the Formula (1) of modularity function *Q*, the function *Q* can be converted as the following:

$$Q_i = e_{ii} - a_i^2 \ , \ Q = \sum_i Q_i \tag{4}$$

Formula (4) suggests that the *Q* value of community detection result is the cumulative result of Q_i value of each community. A higher Q_i value of each community indicates a higher total value of *Q*.

Table 1. Example of One-Way Crossover

V	A(Source)	B(Destination)	B(new)
1	2	6	6
2→	(1)→	2→	1
3	(1)→	2→	1
4	(1)→	2→	1
5	5	2	2
6	8	1	1
7	8	1	1
8	7	6	6

Based on the above analysis, anti-destructive one-way crossover operation is given. The main idea is when source individual A and destination individual B are carried out crossover operation, first node i is chosen arbitrarily in network and its corresponding community label l_i in source individual A also will be obtained. Then, copy individual *B* so as to form individual *B* and individual $B^{destination}$. And assign all the labels whose gene values are l_i in individual A to the community labels of corresponding genes in individual $B^{destination}$ so as to form the new individual B^{new}. Afterwards, the corresponding Q_i values of community whose label is l_i in individual A and B^{new} will be calculated respectively. And they are represented as Q_i^A and $Q_i^{B^{new}}$. If $Q_i^{B^{new}} > Q_i^A$, B^{new} will be reserved and individual *B* will be deleted. Otherwise, B^{new} will be deleted.

It can be seen that anti-destructive one-way crossover operation makes a judgment about the gene value transmitted by individuals during the crossover process of source individual to destination individual whether it will have a positive effect on the community detection result of destination individual. If the transmitted information cannot ensure the newly-generated individuals are better, those individuals will be abandoned, as so to ensure community structure will be developed in a better direction through crossover operation and avoid destroying the existing good community structure.

2.5. Mutation Strategy

Currently, most of the existing mutation strategies of community detection approaches based on genetic algorithm have the defect of weak local search ability. As a result, it is hard to recognize the community structure with high quality [17]. Besides, the traditional community detection approaches based on local optimization also have the defects that community boundary cannot be recognized precisely and that stable community structure cannot be obtained from the different nodes inside the community. Thus, it is hard to transplant the approach based on local optimization to mutation strategy of genetic algorithm. Taking the function *R* as an example proposed in reference [18], this problem is illustrated below. And function *R* is defined as follows:

$$R = \frac{B_{in}}{B_{in} + B_{out}} \tag{5}$$

Here, B_{in} is the number of links connecting nodes inside the community and boundary nodes and B_{out} is the number of links connecting nodes out of the community and boundary nodes.

Figure 2 shows that there are three communities: *A*, *B* and *C*. If function *R* is used to analyze which node of *i* and *j* will be combined by community B in the next step, the following conditions are possible:

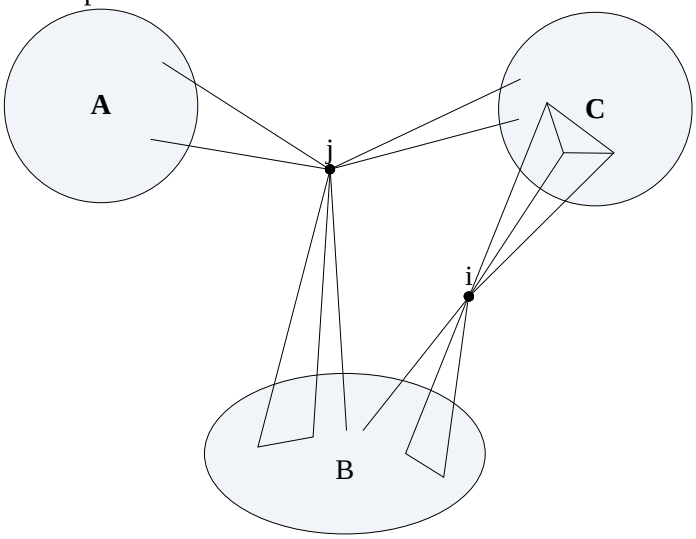

Figure 2. Example of Community Detection of Local Optimization Algorithm

If node *i* is added into community *B*, the value of *R* may be:

$$R_i = \frac{B_{in}+3}{B_{out}-3+3+B_{in}+3} = \frac{B_{in}+3}{B_{out}+B_{in}+3} \tag{6}$$

If node *j* is added into community *B*, the value of *R* may be:

$$R_j = \frac{B_{in}+3}{B_{out}-3+4+B_{in}+3} = \frac{B_{in}+3}{B_{out}+B_{in}+4} \tag{7}$$

Obviously, $R_i > R_j$, in the process of algorithm, node *i* will be firstly integrated into community *B*. Then, according to Figure 2, node *j* has a denser connection with community *B* compared with node *i*. Therefore, node *j* should belong to community *B*. On the contrary, if node *i* belongs to community *C*, many triangles will be formed in community and community connection will be denser. Thus, it is more reasonable that node *i* belongs to community *C*. Through analyzing the reasons of this condition, it can be seen that this method only analyzes the nodes to be combined from intra-community and neglects the view of nodes themselves. Thus, it is hard to precisely detect the community structure. Besides, function *M* proposed in reference [19] and function *L* proposed in reference [20] also have the problem mentioned above, so there is no need to be repeated here. On this basis, considering both the similarity of nodes and intra-community connections and the connection density of nodes and community, the paper proposes the mutation strategy of node local optimization. Before the strategy is illustrated, the definitions are represented as follows:

Definition 1 (Node-Community Similarity): Suppose *t* is a node in network *N* and c_t is the community in connection with node *t*, then $J_t^{c_t}$, the node-community similarity of node *t* and community c_t, can be defined as follows:

$$J_t^{c_t} = \frac{\sum_{r \in c_t \cap adj(t)} \frac{|adj(t) \cap adj(r) \cap c_t|}{|\min(adj(t), adj(r))|}}{|adj(t) \cap c_t|} \tag{8}$$

Here, $adj(t)$ represents the set of neighbor node of node *t*.

It is indicated that definition 1 shows the similarity of node *t* and its neighbor nodes belonging to community c_t. The higher similarity the node *t* and each neighbor node belonging to community c_t have, the higher the node-community similarity $J_t^{c_t}$ is. And it can completely reflect the similarity of nodes and intra-community connections.

Definition 2 (Node Belonging Value) Suppose *t* is a node in network *N and* c_t is a community in connection with node *t*, then $V_t^{c_t}$, the node belonging value of node *t* belonging to community c_t, can be defined as follows:

$$V_t^{c_t} = \frac{|adj(t) \cap c_t|}{|adj(t)|} \tag{9}$$

Obviously, definition 2 shows the connection density of node *t* and community c_t.

Based on the consideration of similarity of nodes and intra-community connections and the connection density of node itself and community, the function of node belonging strength is given.

Definition 3 (The Function of Node Belonging Strength): Suppose $J_t^{c_t}$ is the node-community similarity of node *t* and community c_t and $V_t^{c_t}$ is the node belonging value of node *t* belonging to community c_t, then the node belonging strength Function *F* of node *t* belonging community c_t can be defined as follows:

$$F = \sqrt{J_t^{c_t} \times V_t^{c_t}} \tag{10}$$

It can be known from definition 3 that node belonging strength function considers both the similarity of nodes and intra-community connections and the connection density of node itself and community. It analyzes the community that node should belong to from two aspects of community inside and node itself. On this basis, this paper proposes the mutation strategy of node local optimization. The basic idea is to decide which community the node belongs to by calculating the node belonging strength function of each node. If there is a maximum node belonging strength value between a node and its adjacency community, the label of this node mutates into the community label which has a maximum node belonging strength value. If there are more than one maximum node belonging strength values between node and its adjacency community, the label of this node mutates into the community label which has a maximum node belonging value. The details of mutation strategy of node local optimization are demonstrated in Algorithm 1.

Algorithm1. Mutation strategy of node local optimization

Input: *R* and P_m // *R* is the individual to be mutated and P_m is mutation rate.

Output: R_{new} // R_{new} is the individual after mutation.

Begin

1. for $i = 1{:}n \times P_m$ //*n* is the number of nodes in network and is also the gene number of individual R.

2. $M_{node} \leftarrow$ choose a node to be mutated randomly

3. $adjComs \leftarrow$ obtain the corresponding community labels of all neighbor nodes of node M_{node} to be mutated

4. for each $c \in adjComs$

5. $nbcs(c) \leftarrow$ calculate the node belonging strength of node M_{node} in community c

6. end for

7. $maxnbcs \leftarrow$ get the maximum node belonging strength value in *nbcs*

8. if length(find(*nbcs*==*maxnbcs*))==1

9. $l(M_{node}) \leftarrow$ the corresponding community label of *maxnbcs*

10. else

11. $l(M_{node}) \leftarrow$ the community label of the corresponding community of maximum node belonging value neighboring node M_{node}

12. end if

13.end for

End

Thus, in the process of promoting the mutation of node label, mutation strategy of node local optimization not only considers the similarity of nodes and intra-community connections, but also considers the connection density of nodes and community. Here Figure 2 is the example to be illustrated. It can be seen that in the process of choosing belonging community, the node belonging strength values of node *i* that belong to community *C* and community *B* are respectively $F_i^C \approx 0.577$ and $F_i^B \approx 0.408$. Thus, the label of node *i* should mutate into the label of community *C*. In the process of choosing belonging community, the node belonging strength values of node *j* that belong to community *A*, community *B* and community *C* are respectively $F_j^A = 0$, $F_j^B \approx 0.377$ and $F_j^C = 0$. Thus, the label of node *j* should mutate into the label of community *B*. Therefore, it is reasonable to consider the node belonging community from the two aspects of similarity of node and intra-community connection and the connection density of nodes and community.

2.6. Description of IGALO Algorithm

In genetic algorithm, good population selection strategy is indispensable to reserve good individual. Thus, this paper carries out selection operation referring to the idea of reference [16]. The basic idea is to choose P_{size} individuals of the maximum objective function among parent population and population after crossover and mutation as the child population. P_{size} is the population size of each generation in genetic algorithm.

Based on label propagation method of one-iteration, anti-destructive one-way crossover strategy, mutation strategy of node local optimization and the population selection strategy proposed in reference [16], this paper proposes IGALO algorithm. Details see algorithm 2.

Algorithm 2. IGALO Algorithm

Input: N,P_{size} // N represents complex network and P_{size} is population size.

Output:S // represents the optimal community structure

Begin

1.$Y \leftarrow$ generate initial population in label propagation method of one-iteration

2.tempMaxQ $\leftarrow$ calculate the optimal Q value of individual in Y

3.parentMaxQ $\leftarrow 0$

4.while tempMaxQ > parentMaxQ

5. parentMaxQ $\leftarrow$ *tempMaxQ*

6. $Y_{new} \leftarrow$ carry out anti-destructive one-way crossover operation to individuals in Y

7. $Y_{new} \leftarrow$ carry out mutation operation of node local optimization to individuals in Y_{new}

8 .*tempMaxQ* $\leftarrow$ optimal *Q* value in population Y_{new}

9. $Y_s \leftarrow Y_{new} \cup Y$

10. $Y \leftarrow$ choose P_{size} individuals with optimal *Q* value in Y_s

11.end while

12.$S \leftarrow$ choose optimal individuals in *Y*

End

In IGALO algorithm, the purpose of label propagation strategy of one-iteration is to generate initial population with certain precise, so as to fully realize the searching advantage of genetic algorithm. Anti-destructive one-way crossover strategy is adopted to avoid destroying the original good community structure in the process of crossover and ensure the algorithm to search in the direction of making community structure better. Mutation strategy of node local optimization is adopted to ensure the precision of community detection, considering the node belonging community from two aspects of similarity of nodes and intra-community connections and connection density of nodes and community. The population selection strategy in reference [20] is used to reserve the good individuals in population. Whether the optimal *Q* value of child population after crossover and mutation is greater than the optimal *Q* value of parent population is the cease condition of algorithm, so as to achieve the globally searching ability of algorithm.

2.7. Analysis of Algorithm Time Complexity

Suppose *n* is the number of nodes in network, P_{size} is the population size, L is the times of iteration and that *k* is the average degree of node. The analysis process of IGALO algorithm time complexity is as follows.

Obviously, when population is initialized, the time complexity of label propagation method of one-iteration is $O(P_{size}kn)$. In the operation of anti-destructive one-way crossover, when any two individuals are carried out crossover operation, the time complexity is $O(n)$ and the time complexity of crossover operation to all individuals of population is no more than $O(LP_{size}n)$. In mutation operation of node local optimization, the node belonging community is judged by calculating the node belonging strength. The mutation time complexity of one node is $O(k)$, then the time complexity of corresponding globally mutation operation is $O(LP_{size}kn)$. The time complexity of selecting population is $O(LP_{size}n)$. Above all, the time complexity of IGALO is $O(P_{size}kn + LP_{size}n + LP_{size}kn + LP_{size}n)$.

It is necessary to illustrate that the population size P_{size} and the iterative times *L* are constants, so the time complexity of IGALO algorithm is $O(ckn)$. Complex networks usually are sparse networks and *k* is much less than *n*, so the time complexity of IGALO algorithm is near $O(cn)$, where *c* is a constant.

3. Experiments

To prove the effectiveness of algorithms proposed in this paper, IGALO algorithm is to be tested on the datasets of benchmark networks and real-world networks and compared with GN algorithm based on division, FN algorithm and BGLL algorithm based on local modularity optimization, LPA algorithm based on label propagation and classic GANET algorithm based on genetic algorithm.

Parameter settings are needed when genetic algorithm is running, so IGALO algorithm also needs general parameter settings of genetic algorithm. On the basis of references [9-

15] and analysis of them, this paper gives the parameter settings of IGALO algorithm and the parameter settings are presented in Table 2.

Table 2. IGALO Parameter Settings

Parameter	Meaning	Value
P_{size}	Population size	100
P_c	Crossover rate	70%
P_m	Mutation rate	30%

3.1. Evaluation Criteria

Normalized mutual information (*NMI)* [21] proposed by Danon is the first evaluation criteria chosen by this paper, which is best fit for the accuracy of community detection. *NMI* is used to evaluate the similarity of true community structure and community structure detected by algorithms and the definition can be defined as follows:

$$I(A,B)=\frac{-2\sum_{i=1}^{CA}\sum_{j=1}^{CB}N_{ij}\log(N_{ij}N/N_{i.}N_{.j})}{\sum_{i=1}^{CA}N_{i.}\log(N_{i.}/N)+\sum_{j=1}^{CB}N_{.j}\log(N_{.j}/N)} \tag{11}$$

Here, *N* is matrix. And matrix rows correspond to true community structure and matrix columns correspond to community structure detected by algorithms. N_{ij} is the number of coincided nodes of true community *i* and community *j* detected by algorithms. $N_{i.}$ is the sum of elements of ith row and $N_{.j}$ is the sum of elements of jth column. *CA* represents the number of true communities and *CB* represents the number of communities detected by algorithms. That is to say, the higher accuracy the community structure detected by algorithms has, the greater the value of *NMI* is. Otherwise, the smaller the value of *NMI* is.

The second evaluation criteria adopted by this paper is modularity function *Q* proposed by Newman. It is widely used to evaluate the connection density of communities. The details of modularity function *Q* see Formula (1).

3.2. Dataset Experiment of Benchmark Networks

LFR benchmark [22] network has extremely similar statistical property with real-world networks, so this paper adopts the benchmark network as the experimental dataset of IGALO. The parameter settings are as follows: the network size N is set to 200, the average node's degree k is set to 15, the maximum node's degree is 50, the mixing parameter (the ratio of the link number of each node and outside community to its total node's degree) μ =0.1-0.5. With the increase of μ value, the community structure of networks is fuzzier and fuzzier and great challenges are also brought to algorithms of community detection.

Figure 3 lists the comparison of IGALO and other algorithms in terms of *NMI*. In the comparison with GANET algorithm, when μ =0.1, the *NMI* value of GANET algorithm is greater, but still lower than IGALO algorithm. Then, with the increase of μ value, when the community structure is fuzzier and fuzzier, the NMI value decreases obviously. The reason is that GANET algorithm adopts destructive crossover strategy and mutation strategy, thus searching ability of the algorithm is influenced. In the comparison with LPA algorithm, the NMI value is equal to IGALO algorithm when μ =0.1-0.3. But when the community structure is fuzzier and fuzzier, the accuracy of community detection decreases rapidly. The reason for this is that the algorithm is greatly random. In the

comparison with FN algorithm, because FN algorithm only takes the connection density of intra-community into consideration and neglects the node belonging community from node aspect in the process of community detection, so when μ =0.1-0.5, the NMI value is lower than IGALO algorithm. When μ =0.1-0.3, the NMI value of GN algorithm is higher, but when μ =0.4-0.5, the accuracy of community detection is lower than IGALO algorithm. In the comparison with BGLL algorithm, when μ =0.4, IGALO algorithm has some fluctuations, the NMI value is little lower than BGLL algorithm. But when μ =0.5 and the community structure is very fuzzy, IGALO algorithm can still get higher *NMI* value. It also indicates that IGALO algorithm has stronger community detection ability than other comparing algorithms.

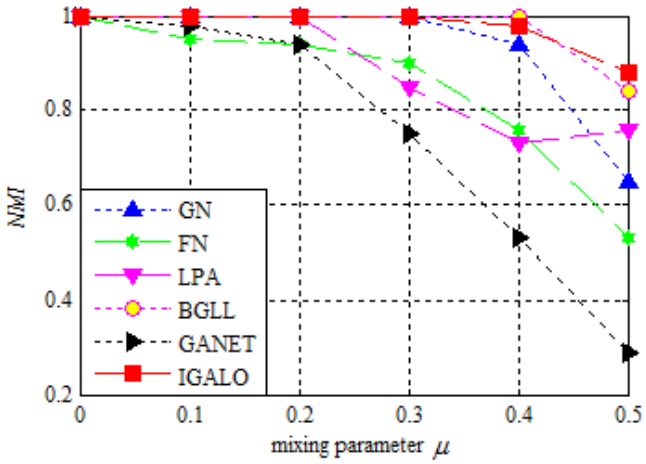

Figure 3. NMI Value Comparison of Different Algorithms

Figure 4, lists the comparison of different algorithms in terms of Modularity function *Q*. From it we can see that Q value of IGALO algorithm when $\mu = 0.1-0.5$ is superior to other algorithms. Thus, IGALO is also able to detect dense community structure within the community.

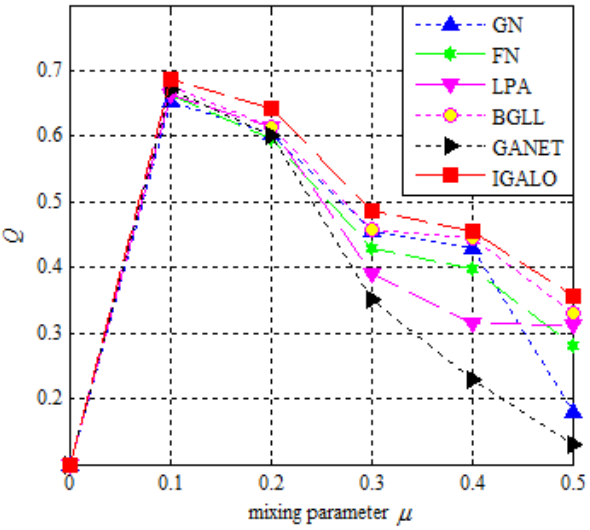

Figure 4. Q Value Comparison of Different Algorithms

3.3. Dataset of Real-World Networks

To further value the detection ability of different algorithms in real-world networks, this section uses the ten real-world networks listed in Table 3 to analyze the different algorithms, including small-size real-world networks with dozens of nodes and big-size real-world networks with tens of thousands of nodes.

Table 3. Datasets of Real Networks

networks	nodes	links	average degree	description
Karate	34	78	4.59	Karate Club network [23]
Dolphins	62	159	5.13	Dolphin social network [24]
Lesmis	77	254	6.6	Lesmis relationship network [25]
Polbooks	105	441	8.4	Books about US politics network [26]
Football	115	613	5.33	American university football club network [2]
Jazz	198	2,742	13.85	Jazz band cooperation network [27]
Email	1,133	5,451	9.62	E-mail communication network [28]
Polblogs	1,490	19,022	25.53	American Presidential Election blog network [29]
Powergrid	4,941	6,594	4.55	Western powergrid of US network [30]
Internet	22,963	48,436	4.22	Internet snapshot network [31]

Among those networks, Karate, Dolphins, Polbooks and Football have true community structure, so the evaluation criteria NMI is adopted to analyze the community detection of different algorithms. Table 4, gives the community detection accuracy of different algorithms in four datasets of known community structure. It is can be seen that the detection accuracy of IGALO is little lower than GANET in Dolphins, while the detection accuracy of IGALO is much higher than other algorithms in other datasets.

Table 4. Comparing the Different Algorithm in Term of NMI on Real-World Networks

Algorithms	Karate	Dolphins	Polbooks	Football
GN	0.5798	0.5542	0.5585	0.8789
LPA	0.8372	0.5914	0.5307	0.8988
FN	0.6925	0. 5727	0.5308	0.7571
BGLL	0.5866	0.5162	0.5745	0.8903
GANET	0.5998	**0.6981**	0.5930	0.6656
IGALO	**1.00**	0.6530	**0.6165**	**0.9268**

In order to further illustrate the effectiveness of IGALO algorithm, community structure detected by IGALO in terms of four real-world networks is to be analyzed.

Karate is an American karate club network, which was set up after two-year observation and analysis about an American karate club. 34 members of the club are regarded as nodes and the friendship of members is regarded as the links connecting two nodes. Due to dispute, the club is finally split into two new clubs which are respectively president-centered (34 nodes) and coach-centered (one node). The NMI value that IGALO gets in terms of Karate is 1 and the corresponding community structure is showed in Figure 5. It can be seen that IGALO algorithm accurately divides karate network into two communities.

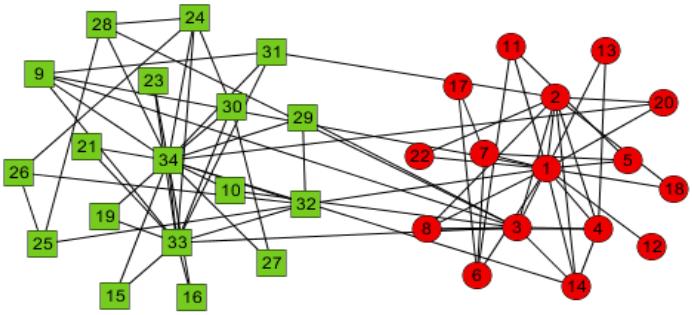

Figure 5. Community Structure Detected by IGALO in Terms of Karate Network

Dolphins was set up to observe 62 dolphins living in Doubtful Sound of New Zealand. Every dolphin represents a vertex. If two dolphins often communicate with each other, and there will be a link between them. The *NMI* value IGALO gets in terms of Dolphins is 0.6530 and the corresponding community structure is given in Figure 6. From Figure 6, we can see that IGALO divides Dolphins into four communities. And the part which is represented by a square is the same as the true community structure of Dolphins. And the other part of true community structure is further divided into smaller and denser communities represented by circulars, triangles and rhombuses.

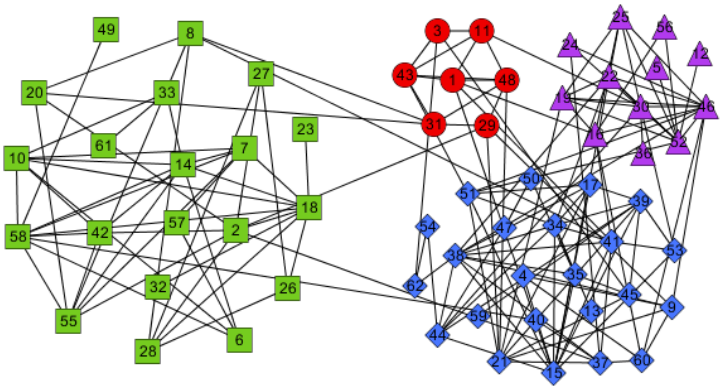

Figure 6. Community Structure Detected by IGALO in Terms of Dolphins Network

Polbooks network is books about US politics network. Nodes mean selling books about US politics on Amazon.com and there are 105 books. Links mean books that the same customer bought and there are 441 links totally. The political views that those book represent include ideas of Progressives, Centrists and Conservatives. In terms of Polbooks, *NMI* value that IGALO algorithm gets is 0.6165 and the corresponding community structure is showed in Figure 7. It can be seen that IGALO algorithm divides it into four communities. And the books of Progressives are divided into two smaller communities respectively represented by rhombus and triangle. Because the political views of Centrist books are vague and readers frequently buy the books with other two views, the Centrist books are divided into other groups. According to the opinions of reference [26], this division is relatively logical.

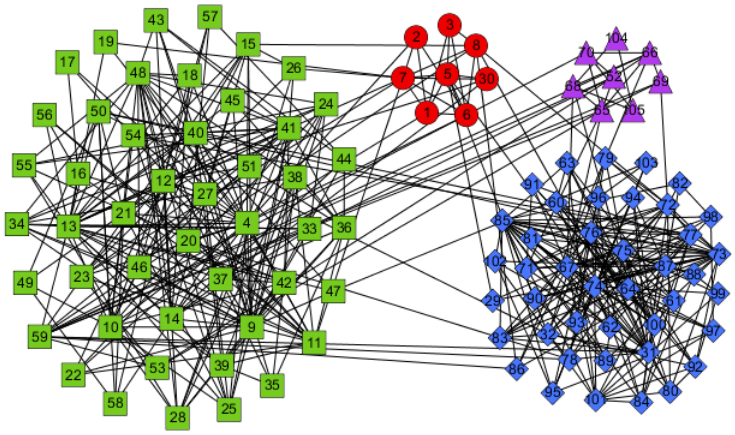

Figure 7. Community Structure Detected by IGALO in Terms of Polbooks Network

Football network is American university football club network, which was set up according to the game plan of 2000 season. And it is divided into 12 associations. Nodes represent the competing teams and there are 115 nodes totally. Links mean there are competitions between two teams in this season and there are 613 links totally. The *NMI* value of community detected by IGALO algorithm is 0.9268 and the corresponding community structure is showed in figure 8. It can be seen that except that few nodes are misplaced, IGALO algorithm almost perfectly divides Football network into 12 communities.

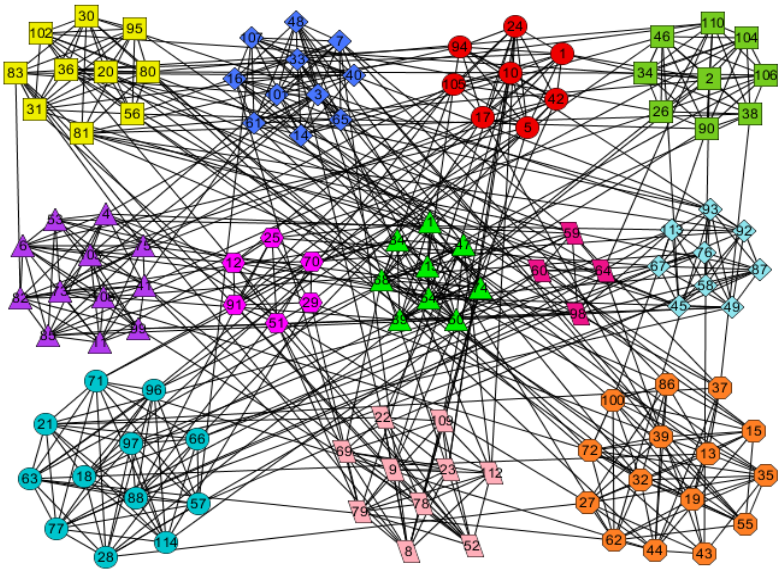

Figure 8. Community Structure Detected by IGALO in Terms of Football Network

Currently, as the evaluation criteria of community structure, modularity function *Q* has been widely accepted by researchers. Thus, the evaluation criteria, function *Q* and the testing dataset, ten real-world networks listed in Table 3, are used to comparatively analyze IGALO algorithm and other algorithms. Table 5, lists the *Q* value of different algorithms on real-world networks. "\" means algorithm fails to detect community or it fails to detect community structure within 48 hours.

It can be seen that IGALO algorithm is optimal on 6 networks among the ten networks and four networks rank the second. *Q* value of IGALO algorithm is lower than BGLL algorithm only on Karate and Email and is lower than FN algorithm on Powergrid and Internet. GN algorithm needs global analysis to get the number of edge betweenness and finishes community detection through splitting, so the efficiency of the algorithm is very low. Thus, on the network with more than 1000 nodes, GN algorithm cannot complete the task of community detection. While LPA algorithm has strong randomness, so the community detection quality is influenced. The corresponding *Q* value of community structure detected by FN and BGLL on some networks is superior to IGALO, but in general IGALO is superior to the other two algorithms. GANET algorithm based on genetic algorithm adopts destructive crossover strategy and mutation strategy with strong randomness in the process of community detection, so the overall detection quality of GANET algorithm is not so high. Through the comparative analysis of *Q* value with different algorithms on ten real-world networks, the community detection ability of IGALO algorithm is better.

Table 5. Comparing the Different Algorithms in Terms of Q Value of Ten Real-World Networks

Networks	GN	LPA	FN	BGLL	GANET	IGALO
Karate	0.4013	0.4020	0.3807	**0.4188**	0.3998	0.4172
Dolphins	0.5194	0.5113	0.4897	0.5188	0.4824	**0.5268**
Lesmis	0.5381	0.5289	0.5006	0.5556	0.4829	**0.5571**
Polbooks	0.5168	0.5156	0.5020	0.4986	0.4906	**0.5264**
Football	0.5996	0.5977	0.5773	0.6021	0.5840	**0.6046**
Jazz	0.4051	0.4391	0.3103	0.4431	0.2568	**0.4435**
Email	\	0.4943	0.5037	**0.5406**	0.2729	0.5149
polblogs	\	0.4262	0.4269	0.4251	0.1896	**0.4356**
PowerGrid	\	0.6114	**0.9341**	0.7490	0.6815	0.7803
Internet	\	0.4989	**0.6379**	0.4592	0.4790	0.5334

4. Conclusion

This paper comes up with the community detection in complex networks based on improved genetic algorithm and local optimization (IGALO) in terms of the defect that traditional community detection approaches based on genetic algorithm have strong randomness and weak searching ability in the process of community detection. Taking modularity function *Q* as the objective function, IGALO algorithm adopts string encoding method as encoding method of algorithm. Label propagation method of one-iteration is used to initialize population so as to generate initial population with certain precision and to fully realize the searching ability of genetic algorithm. In terms of the defect that traditional crossover strategy easily destroys the existing good community structure, anti-destructive one-way crossover strategy is proposed to ensure the algorithm to search in the direction of increasing *Q* value. In terms of the problem that the searching ability of existing community detection approaches based on genetic algorithm is weak, mutation strategy of node local optimization is proposed, so in the process of node mutation, not only the connection density of nodes and community but also the similarity of nodes and intra-community connections will be considered, so as to ensure the accuracy of node belonging community and improve the quality of community detection. It can be seen that IGALO algorithm aims to decrease the randomness of algorithm in every stage of searching good community structure, so as to ensure the quality of community detection. Tests are made on benchmark networks and real-world networks and comparative analysis is also made with various classic algorithms. The results show that IGALO

algorithm can get better community structure both on small-size networks with dozens of nodes and big-size networks with tens of thousands of nodes, so IGALO is proved to be effective and feasible.

Acknowledgments

This work is sponsored by the Humanity and Social Science Youth foundation of Ministry of Education of China under Grant No. 15YJCZH088.

References

[1] M. E. J. Newman and M Girvan, "Finding and evaluating community structure in networks", Physical Review E, vol. 69, no. 2, **(2004)**, pp. 026113.
[2] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 12, **(2002)**, pp. 7821-7826.
[3] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", Nature, vol. 435, no. 7043, **(2005)**, pp. 814-818.
[4] Y. Y. Ahn, J. P. Bagrow and S. Lehmann, "Link com munities reveal multiscale complexity in networks", Nature, vol. 466, no. 7307, **(2010)**, pp. 761-764.
[5] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical Review E, vol. 76, no. 3, **(2007)**, pp. 036106.
[6] A. Lancichinetti, S. Fortunato and J. Kertesz, "Detecting the overlapping and hierarchical community structure in complex networks", New Journal of Physics, vol. 11, **(2009)**, pp. 033015.
[7] V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks", Journal of Statistical Mechanics:Theory and Experiment, vol. 2008, no. 10, **(2008)**, pp. P10008.
[8] Y. Q. Hu, M. H. Li, P. Zhang and Z. Di, "Community detection by signaling on complex networks", Physical Review E, vol. 78, no. 1, **(2008)**, pp. 016115.
[9] 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, Germany, **(2008)** September 13-17.
[10] M. Tasgin, A. Herdagdelen and H. Bingol, "Community detection in complex networks using genetic algorithms", arXiv preprint arXiv, **(2007)**, pp. 0711.0491.
[11] C. Shi, Z. Yan, Y. Wang, Y. N. Cai and B. Wu, "genetic algorithm for detecting communities in large-scale complex networks", Advances in Complex Systems, vol. 13, no. 1, **(2010)**, pp. 3-17.
[12] D. X. He, X. Zhou, Z. Wang, C. G. Zhou, Z. Wang and D. Jin, "Community mining in complex networks-clustering combination based genetic algorithm", Acta Automatica Sinca, vol. 36, no. 8, **(2010)**, pp. 1160-1170.
[13] M. G. Gong, B. Fu, L. C. Jiao and H. Du, "Memetic algorithm for community detection in networks", Physical Review E, vol. 84, no. 5, **(2011)**, pp. 056101.
[14] D. Y. Liu, D. Jin, C. Baquero, D. X. He, B. Yang and Q. Y. Yu, "Genetic algorithm with a local search strategy for discovering communities in complex networks", International Journal of Computational Intelligence Systems, vol. 6, no. 2, **(2013)**, pp. 354-369.
[15] R. H. Shang, J. Bai, L. C. Jiao and C. Jin, "Community detection based on modularity and an improved genetic algorithm", Physica A: Statistical Mechanics and its Applications, vol. 392, no. 2, **(2013)**, pp. 1215–1231.
[16] E. Mezura-Montes and C. A. C. Coello, "A simple multimemberd evolution strategy to solve constrained optimization problems", IEEE Transactions on Evolutionary Computation, vol. 9, no. 1, **(2005)**, pp. 1-17.
[17] D. Jin, J. Liu, B. Yang, D. X. He and D. Y. Liu, "Genetic algorithm with local search for community detection in large-scale complex networks", Acta Auto-matica Sinica, vol. 37, no. 7, **(2011)**, pp. 873-882.
[18] A. Clauset, "Finding local community structure in networks", Physical Review E, vol. 72, no. 2, **(2005)**, pp. 026132.
[19] F. Luo, J. Z. Wang and E. Promislow, "Exploring local community structures in large networks", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Hong Kong,China, **(2006)** December 18-22.
[20] J. Y. Chen, O. R. Zaiane and R. Goebel, "Local community identification in social networks", Proceedings of the 2009 Internation Conference on Advances in Social Network Analysis and Mining, Athens, Greek, **(2009)** May 26-29.
[21] L. Danon, A. D. Guilera, J. Duch and A. Arenas, "Comparing community structure identification" , Journal of Statistical Mechanics:Theory and Experiment, **(2005)**, pp. P09008.

[22] A Lancichinetti, S Fortunato, F Radicchi, "Benchmark graphs for testing community detection algorithms" ,Physical Review E, vol. 78,no. 4,**(2008)**, pp. 046110.
[23] W. W. Zachary, "An information flow model for conflict and fission in small groups" ,Journal of Anthropological Research, vol. 33,no. 4, **(1977)**, pp. 452-473.
[24] D. Lusseau, "The emergent properties of a dolphin social network" ,Proceedings of the Royal Society B: Biological Sciences, vol. 270, no. S2, **(2003)**, pp. 186-188.
[25] D. E. Knuth, "The Stanford GraphBase:A Platform for Combinatorial Computing" , http://www-cs-faculty.stanford.edu/~uno/sgb.html, **(2015)**.
[26] M. E. J. Newman, "Modularity and community structure in networks", Proceedings of the National Academy of Sciences of the United States of America, vol.103, no. 23, **(2006)**, pp. 8577–8582.
[27] P. M. Gleiser and L. Danon, "Community structure in jazz",Advances in Complex Systems,vol. 6, no. 4,**(2003)**, pp. 565-573.
[28] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt and A .Arenas, "Self-similar community structure in a network of human interactions", Physical review E, vol. 68, no. 6, **(2003)**, pp. 065103.
[29] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog", Proceedings of the 3rd International Workshop on the Weblogging Ecosystem, New York, USA, **(2005)**, pp. 36-43.
[30] D. J Watts and S. H. Strogatz, "Collective dynamics of 'small- world' networks", Nature, vol. 393, no. 84, **(1998)**, pp. 440-442.
[31] M. E. J. Newman, "Network data from mark newman's home page" , http://www-personal.umich.edu/~mejn/netdata , **(2015)**.

