A Non-dominated Neighbor Immune Algorithm for Community Detection in Networks

Maoguo Gong
Xidian University
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China,
Xi'an, 710071, China
gong@ieee.org

Tian Hou, Bao Fu
Xidian University
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China,
Xi'an, 710071, China
{houtian608, fubao2007}@gmail.com

Licheng Jiao
Xidian University
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China,
Xi'an, 710071, China
lchjiao@ieee.org

ABSTRACT

The study of complex networks has received an enormous amount of attention from the scientific community in recent years. In this paper, we propose a multi-objective approach, named NNIA-Net, to discover communities in networks by employing Non-dominated Neighbor Immune Algorithm (NNIA). Our algorithm optimizes two objectives to find communities in networks— groups of vertices within which connections are dense, but between which connections are sparser. The method can produce a series of solutions which represent various divisions to the networks at different hierarchical levels. The number of subdivisions is automatically determined by the non-dominated individuals resulting from our algorithm. We demonstrate that our algorithm is highly efficient at discovering quality community structure in both synthetic and real-world network data. What's more, a new initialization method is proposed to improve the traditional initialization method by about 30% in running time.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search-*Heuristic Methods*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Multi-objective optimization, artificial immune system, non-dominated solutions, community detection, complex networks.

1. INTRODUCTION

Recent research indicates that a large body of diverse systems in many different domains such as collaboration networks, the Internet, the World Wide Web, biological networks, communication and transport networks, social networks and so on can be represented as complex networks [2].

In the continuing flurry of research activity within physics and mathematics on the properties of complex networks, a particular recent focus has been the analysis of communities within networks [11]. Real networks are not random graphs, and they reveal a high level of order and organization. The degree distribution is broad, with a tail that often follows a power law [12]: therefore, many vertices with low degree coexist with some vertices with large degree. Furthermore, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups. This feature of real networks is called community structure, or clustering. Communities, also called clusters or modules probably share common properties and/or play similar roles within the graph [12]. They are groups of vertices having denser connections within them, and sparser connections between the groups. In protein-protein interaction networks, communities are likely to group proteins having the same specific function within the cells, in the graph of the World Wide Web, they may correspond to groups of pages dealing with the same or related topics, in metabolic networks, they may be related to functional modules such as cycles and pathways, in food webs, they may identify compartments, and so on.

1.1 Related Work

Community structure is an important network property and can reveal many hidden features of the given networks. Hence, community identification is a fundamental step for discovering what makes entities come together, but also for understanding the overall structural and functional properties of large network [13]. The capability of detecting the partitioning of a network in clusters can give important information and useful insights to understand how the structure of ties affects individuals and their relationships. The problem of community detection has been receiving a lot of attention and many different approaches have been proposed.

GA-Net [3], a Genetic Algorithm for community detection in social networks proposed by Clara Pizzuti in 2008 introduces the concept of community score to measure the quality of a partitioning of a network in communities, and tries to optimize this quantity by running the Genetic Algorithm. All the dense communities present in the network structure are obtained at the end of the algorithm by selectively exploring the search space, without the need to know in advance the exact number of groups [3]. In GA-Net, only one objective function, community score is optimized, so that only a certain solution is obtained in one run. Unlike many existing methods, the algorithm does not require the number of communities to find. This number is automatically determined by the optimal value of the community score [3].

GECCO'11, July 12–16, 2011, Dublin, Ireland.

Despite all the advantages, the drawbacks of GA-Net can not be neglected. Only one objective function is not enough to determine whether a partitioning is the best one, and just one possible division to a network provided in one run is far away from the need of people. Because of that, Clara Pizzuti provided MOGA-Net [4] employing Multi-Objective Genetic Algorithm to uncover community in complex networks. This algorithm introduces two objective functions. The first objective function employs the concept of community score to measure the quality of the division in communities of a network. The higher the community score, the denser the clustering obtained [4]. The second defines the concept of fitness of the nodes belonging to a module and iteratively find modules having the highest sum of node fitness, in the following referred as community fitness. When this sum reaches its maximum value, the number of external links in minimized. Both the objective functions have a positive real-valued parameter controlling the size of the communities. The higher the value of the parameter, the smaller the size of the communities found. MOGA-Net exploits the benefits of these two functions and obtains the communities present in the network by selectively exploring the search space, without the need to know in advance the exact number of groups [4]. This number is automatically determined by the optimal compromise values of the objectives. An interesting result of the multi-objective approach is that it returns not a single partitioning of the network, but a set of solutions [4]. Each of these solutions corresponds to a different trade off between the two objectives and thus to diverse partitioning of the network consisting of various number of clusters. This gives the readers a great chance to analyze several clustering at different hierarchical levels [4].

Community score and community fitness are the two objective functions used in this paper. A fundamental measure criterion modularity Q which will be introduced in detail later used as an evaluation metrics in our paper has been also widely used recently [1, 2, 3]. Modularity Q was used in [14] which optimized network modularity using genetic algorithm to detect community. It is scalable to very large networks and does not need any priori knowledge about the number of communities or any threshold value.

However, Fortunato and Barthélemy[7] showed mathematically that the optimization of modularity has a resolution limit, raising important concerns about the reliability of the modules detected so far using this technique, or eventually using any other quality function [17]. Recently, to conquer this drawback, series of algorithms has been proposed. In [10], the authors provided a method that allows the full screening of the topological structure at any resolution level using the original definition of Q. In [16], the authors presented a mathematical formulation of influence, defined an influence-based modularity metric, and used it to partition the network into communities. And this algorithm outperformed the edge-based modularity algorithm on the standard data sets used in literature. In [6], to deal with the community structure detection problem in directed networks, E .A. Leicht and M. E. J. Newman generalized the widely used benefit function known as modularity in a principled fashion to incorporate the information contained in edge directions.

1.2 Our Main Contribution

In this paper, we propose a multi-objective approach, named NNIA-Net, to uncover communities in networks by employing Non-dominated Neighbor Immune Algorithm (NNIA). Non-dominated Neighbor Immune Algorithm (NNIA) proposed by Gong et al. in 2007 is an algorithm for multi-objective optimization by using a novel non-dominated neighbor-based selection technique, an immune inspired operator, two heuristic search operators, and elitism [15]. NNIA-Net optimizes two objective functions introduced in [3] and [1]. These two objective functions are also used in MOGA-Net [4] which is mentioned above. In NNIA-Net same as in NNIA, only partial non-dominated individuals with greater crowding-distance values are selected to do proportional cloning, uniform crossover and mutation. Because of that, in a single generation, NNIA-Net only pays more attention to the less-crowded regions in the current trade-off front. The uniform crossover operator and mutation operator we used in this paper were described in [3, 4] and Normalized Mutual Information (NMI) and modularity Q are used as the evaluate metrics.

The remainder of this paper is organized as follows: In the next section a description of the problem we deal with and two modularity measures, community score and community fitness designed to evaluate whether a particular partition of a network is good or not, and the multi-objective optimization algorithms with NNIA in detail are given. In section 3, we describe the outline of and each operation introduced in our algorithm in detail. In section 4, we give a number of applications of our algorithms to particular networks, both synthetic networks and real social networks compared with GN. In section 5, we give our conclusions.

2. BACKGROUNDAND

2.1 Problem Statement

The problem we deal with is how to detect community structure in networks. As referred in [4], both synthetic networks and real life networks N can be modeled as a graph G =(V;E) where V is a set of objects, called nodes or vertices, and E is a set of links, called edges, that connect two elements of V. A community is generally thought of as a part of a network where internal connections are denser than external ones. This definition of community is rather vague and there is no general agreement on the concept of density [4]. To sharpen the use of detection algorithms a more precise definition is needed. Many possible definitions of communities exist in the literature [18]. A more formal definition has been introduced in [8].

2.1.1 Related Terms and Definitions

The adjacency matrix A, fully specifies the topology of the network. In the simplest case of an unweighted, undirected network, it is equal to 1 if i and j are directly connected; it is equal to zero otherwise.

The degree k_i of a generic node i, defined as: $k_i = \sum_j A_{ij}$.

Let $S \subset G$ the sub-graph where node i belongs to, the degree of i with respect to S, defined as: $k_i(S) = k_i^{in}(S) + k_i^{out}(S)$.

Where $k_i^{in}(S) = \sum_{j \in S} A_{ij}$ is the number of edges connecting i to the other nodes in S, and $k_i^{out}(S) = \sum_{j \notin S} A_{ij}$ is the number of edges connecting i to the rest of the network.

2.2 Measures of a Partitioning to a Graph

The two objective functions are introduced as follows:

2.2.1 Community Score

In [3, 4], the concept of community score of a graph is defined as:

Let μ_i denote the fraction of edges connecting node i to the other nodes in S. It is defined as:

$$\mu_i = \frac{1}{|S|} k_i^{in}(S) \tag{1}$$

where |S| is the cardinality of S.

The power mean of S of order r, denoted as M(S) is defined:

$$M(S) = \frac{\sum_{i \in S} (\mu_i)^r}{|S|} \tag{2}$$

Notice that, in the computation of M(S), since $0 \le \mu \le 1$, the exponent r increases the weight of nodes having many connections with other nodes belonging to the same module, and diminishes the weight of those nodes having few connections inside S.

The volume v_S of a community S is defined as the number of edges connecting vertices inside S, i.e. the number of 1 entries in the adjacency sub-matrix of A corresponding to S, v_S is defined as:

$$v_S = \sum_{i,j \in S} A_{ij} \tag{3}$$

The score of S is defined as:

$$score(S) = M(S) \times v_S \tag{4}$$

The community score of a partitioning $\{S_1, \cdots S_k\}$ of a graph G is defined as:

$$CS = \sum_{i=1}^{k} score(S_i) \tag{5}$$

The community score gives a global measure of the network division in communities by summing up the local score of each module ($score(S_i)$) found. The problem of community identification has been formulated in [3] as the problem of maximizing CS.

2.2.2 Community Fitness

As referred in [1], communities are essentially local structures, involving the nodes belonging to the modules themselves plus at most an extended neighborhood of them. Social communities are local structures without any reference to the humankind as a whole. The authors of [1] firstly provided the concept of the community fitness of S.

The community fitness of S is defined as:

$$CF(S) = \sum_{i \in S} \frac{k^{in}(S)}{(k^{in}(S) + k^{out}(S))^{\alpha}} \tag{6}$$

where $k^{in}(S)$ and $k^{out}(S)$ are the internal and external degrees of the nodes belonging to the community S, and α is a positive real-valued parameter controlling the size of the communities. The problem of community identification has been formulated in [4] as the problem of maximizing CF.

2.3 Non-dominated Neighbor Immune Algorithm (NNIA)

In this part, we describe a novel multi-objective optimization algorithm, the NNIA[15]. NNIA stores non-dominated individuals found so far in an external population, called the dominant population. Only partial less-crowded non-dominated individuals, called active antibodies, are selected to do proportional cloning, uniform crossover and mutation. Furthermore, the population storing clones is called the clone population. The dominant population, active population, and clone population at time t are represented by time-dependent variable matrices D_t, A_t and C_t, respectively. The main loop of NNIA is as follows [15].

Algorithm 1: Nondominated Neighbor Immune Algorithm (NNIA) [15]

```
Input:
    $G_{\max}$   (maximum number of generations)
    $n_D$   (maximum size of dominant population)
    $n_A$   (maximum size of active population)
    $n_C$   (size of clone population)

Step1: Initialization: Generate an initial antibody population $B_0$ with size $n_D$. Create the initial $D_0 = \phi$, $A_0 = \phi$, and $C_0 = \phi$. Set $t = 0$.
Step2: Update Dominant Population: Identify dominant antibodies in $B_t$. Copy all the dominant antibodies to form the temporary dominant population (denoted by $DT_{t+1}$). If the size of $DT_{t+1}$ is not greater than $n_D$, let $D_{t+1} = DT_{t+1}$. Otherwise, calculate the crowding-distance values of all individuals in $DT_{t+1}$, sort them in descending order of crowding-distance, and choose the first $n_D$ individuals to form $D_{t+1}$.
Step3: Termination: If $t \ge G_{\max}$ is satisfied, export $D_{t+1}$ as the output of the algorithm, Stop; Otherwise, $t = t + 1$.
```

Step4: Non-dominated Neighbor−Based Selection: If the size of D_t is not greater than n_A , let $A_t = D_t$. Otherwise, calculate the crowding-distance values of all individuals in D_t , sort them in descending order of crowding-distance, and choose the first n_A individuals to form A_t .

Step5: Proportional Cloning: Get the clone population C_t by applying proportional cloning to A_t .

Step6: Recombination and Hyper-mutation: Perform recombinetion and hyper-mutation on C_t and set $C_t^{'}$ to the resulting population.

Step7: Get the antibody population B_t by combining the $C_t^{'}$ and D_t ; go to Step2.

Return: $D_{G_{\max+1}}$ (final approximate Pareto-optimal set)

When the number of dominant antibodies is greater than the maximum limitation and the size of dominant population is greater than the maximum size of active population, both the reduction of dominant population and the selection of active antibodies use the crowding-distance based truncation selection. The proportional cloning, uniform crossover, and mutation operators are described in detail in Section 4.

3. ALGORITHM DESCRIPTION

In this section we give a description of the multi-objective algorithm NNIA-Net, employing NNIA [15] to discover community structures in networks.

Our algorithm uses the locus-based adjacency representation proposed in [19], employed by [9] for multi-objective clustering and also employed by [3, 4] for community structure detection in networks. In this representation, every individual of the population consists of N genes $g_1, \ldots\ldots g_N$ and each genetic position can assume allele values j in the range {1,..., N} , where the N is the total number of nodes in this graph. If the value at ith genetic position is j, it means there is a link between ith node with the jth node. And an individual forms a sub-graph with every node connected to certain nodes. The decoding step of this method can been found in [9].

3.1 Initialization Methods

Our initialization process takes in account the effective connections of the nodes in the network. A random generation of individuals could generate components that in the original graph are disconnected. In fact, a randomly generated individual could contain an allele value j in the ith position, but no connection exists between the two nodes i and j, i.e. the edge (i, j) is not present. In such a case it is obvious that grouping in the same cluster both nodes i and j is a wrong choice.

3.1.1 The Traditional Initialization Method

In order to overcome this drawback, referred in [3, 4], a "repair operation" is proposed, once an individual is generated, it is repaired, that is a check is executed to verify that an effective link exists between a gene at position i and the allele value j. This value is maintained only if the edge (i, j) exists. Otherwise, j is substituted with one of the neighbors of i. This guided initialization biases the algorithm towards a decomposition of the network in connected groups of nodes. We call an individual generating this kind of partitioning safe because it avoids uninteresting divisions containing unconnected nodes. Safe individuals improve the convergence of the method because the space of the possible solutions is restricted.

3.1.2 Our New Initialization Method

In this paper, we propose a new approach to generate safe individuals.

3.1.2.1 Basic Concepts

Our new approach introduces two concepts, the first is link table of a graph, second is degree of a node.

The link table of a graph is the ith row of which is the nodes connected to the ith node. So, the link table has N rows, N is the total number of the nodes in a graph. Assume A is the adjacency matrix of a graph, so the ith row of the link table of it is the column number where A (i, :) is equal to 1. For example, if the ith row of A is [1 0 1 0 1 1 1 1 0 0], the ith row of the link table of this graph is [1 3 5 6 7 8 0 0 0 0]. The degree k_i of a generic node i, defined as $k_i = \sum_j A_{ij}$, where A (N*N) is the adjacency matrix of a graph G with N nodes.

3.1.2.2 The Main Steps of Our Initialization Method

Now, we introduce how to produce safe individuals employing the concepts of link table and degree. The nodes connected to the ith node is saved in the ith row of the link table, so when we produce an individual, at the position i, the allele value must be selected in the ith row of the link table of the graph. This method assures every link is effective, this means every link is existed in the origin graph. The safe individuals make sure that there is not uninteresting divisions containing unconnected nodes.

The main steps to generate safe individuals in our new method are as follows:

Algorithm 1: Generate safe individuals

Input: A (adjacency matrix of a graph)
num_nodes (number of nodes)
popsize (size of initial population)

Step1: Create the link table and degree of the node based on the adjacency matrix A of a graph:

Procedure LinkTable

function [link_table,degree] = **LinkTable**(A,num_nodes)

```
link_table = : zeros(num_nodes);
degree = : zeros(num_nodes,1);
for i = 1:num_nodes
    k = 1;
    for j = 1:num_nodes
        if (A(i,j) == 1)
            link_table (i,k) =: link_table (i,k) + j;
            k = k + 1;
        end
    end
    degree (i,1) = k - 1;
end
```

Step2: Create safe population based on link table and degree generated in Step 1:

Procedure InitializePop

function safe_population=**InitializePop**(popsize, num_nodes, link_table, degree)

```
safe_population = zeros(popsize,num_nodes);
for i = 1:popsize
    for j = 1:num_nodes
        safe_population(i,j)=
          link_table(j,ceil(rand*degree(j,1)));
    end
end
```

Return: safe_population

3.1.2.3 The Advantages of Our New Method

The main differences of these two methods are as follows:

1) Traditional method generates a random individual firstly and then repairs it. Once an individual is generated, it is repaired, that is a check is executed to verify that an effective link exists between a gene at position i and the allele value j. This value is maintained only if the edge (i, j) exists. Otherwise, j is substituted with one of the neighbors of i.

2) Our new method firstly finds the available values of each position, and save the values in the link table of this node. Then, when an individual is generated, the value in a certain position only can be chosen from the values in the link table. This operation makes sure every connection is existed in the graph.

3) Without the repair operation, our new method needs less time to generate a population in the same size compared with the method in [5, 6].

In order to compare our new method to generate safe individuals with the method used in [5, 6], simple experiment is operated in Section 5(1). We generate the same size population with the same adjacency matrix in these two different methods, then we compare the times they use separately. The experimental results show that our new method is much more efficient.

3.2 Operators in Our Algorithm

3.2.1 The Cloning Operator

The cloning operator we used in our algorithm is proportional cloning which was introduced in [13] in detail. In immunology, cloning means asexual propagation so that a group of identical cells can be descended from a single common ancestor, such as a bacterial colony whose members arise from a single original cell as the result of mitosis. In this paper, the individual with greater crowding-distance value is reproduced more times. The crowding-distance [19] of a dominant antibody $d \in D$ is given by

$$\zeta(d,D) \triangleq \sum_{i=1}^{k} \frac{\zeta_i(d,D)}{f_i^{\max} - f_i^{\min}} \tag{7}$$

Where $f_i^{\max}$ and $f_i^{\min}$ are the maximum and minimum value of the ith objective and

$$\zeta_i(d,D) = \begin{cases} \infty, & if \quad f_i(d) = m \quad or \quad M \\ \min\{f_i(d^{'}) - f_i(d^{''})\}, & others \end{cases} \tag{8}$$

Where $d^{'}, d^{''} \in D : f_i(d^{''}) < f_i(d) < f_i(d^{'})$,

$m = \min\{f_i(d^{'}) \mid d^{'} \in D\}$, $M = \max\{f_i(d^{'}) \mid d^{'} \in D\}$.

Based on the crowding-distance $\zeta(d,D)$, the density of dominant antibodies surrounding d in the population D can be estimated. The individual with greater crowding-distance value has a larger cloning scale. After we select the active population from the dominated population, each active antibody has a cloning scale according to its crowding-distance. It is mentioned in [13] that how the crowding-distance values of boundary solutions are set. The aim is that the greater the crowding-distance value of an individual, the more times the individual will be reproduced. So there exist more chances to search in less-crowded regions of the trade-off front.

3.2.2 The Crossover Operator

The crossover operator we used in our algorithm is uniform crossover which is suitable to use in the locus-based representation because it guarantees the maintenance of the effective connections of the nodes in the network in the child individual. In fact, because of the biased initialization, each individual in the population is safe, that is it has the property, that if a gene i contains a value j, the n the edge (i, j) exists [3]. Thus, given two safe parents, a random binary vector is created. Uniform crossover then selects the genes where the vector is all from the first parent, and the genes where the vector is a 0 from the second parent, and combines the genes to form the child. The child at each position i contains a value j coming from one of the two parents [4]. Thus the edge (i, j) exists. This implies that from two safe parents a safe child is generated.

3.2.3 The Mutation Operator

The mutation operator that randomly changes the value j of an ith gene causes a useless exploration of the search space, because of possible values an allele can assume are restricted to the neighbor of gene i. This repaired mutation guarantees the generation of a safe mutated child in which each node is linked only with one of its neighbors.

4. EXPERIMENTAL REASULTS

4.1 Our New Initialization Method

In this part, we study the efficiency of our new approach to initialize population on a synthetic data set.

4.1.1 The Benchmark

The synthetic data set we used here is the benchmark proposed by M. Girvan et al. [12] which is a large set of artificial, computer-generated graphs. Each graph is constructed with 128 vertices, each of which is connected to exactly 16 other vertices. The vertices are divided into four separate communities with some number internal degree of each vertex's 16 connections made to randomly chosen members of its own community and the remaining external degree made to random members of other communities. This produces graphs which have known community structure, but which are essentially random in other respects.

Table 1. Running time of the two different methods and the improvement of our method when external degree of a node is 2

popsize	**200**	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**	**1800**
Traditional Method(time1)	0.109	0.203	0.297	0.390	0.485	0.594	0.703	0.781	0.859
Our New Method(time2)	0.063	0.125	0.203	0.266	0.344	0.421	0.485	0.547	0.625
time1-time2	0.046	0.078	0.094	0.124	0.141	0.173	0.218	0.234	0.234
(time1-time2)./time1	0.424	0.384	0.317	0.318	0.291	0.291	0.310	0.300	0.272
Average Improvement	**32.3%**								

4.1.2 The Experimental Results on Our New Initialization Method

Using these graphs, we tested the performance of our new initialization approach and the traditional approach referred in [3, 4]. The running times to generate the same size of population employing the different methods are compared here.

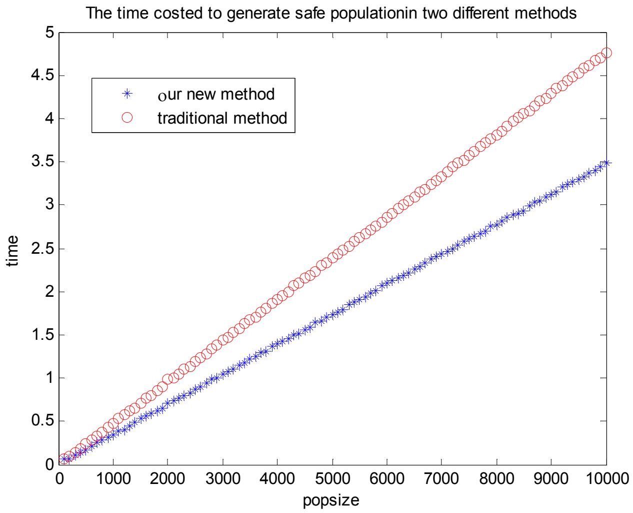


Figure 1. The running time versus the increasing of popsize when we generate safe individuals by employing the two different methods.

Figure 1 shows that in both methods, the running times are increasing with the popsize and our method is much more efficient than the traditional method. The running time of generating 10000 individuals in our method is almost 3.25s and 4.75s in the traditional method. It is nearly improved by 30%. So, larger the size of population is, the more time our method saves.

Table 2. Running time versus the external degree of a node

External Degree	**0**	**2**	**4**	**6**	**8**
t1(s)	48.14	48.03	47.98	47.87	48.12
t2(s)	34.68	34.96	34.71	34.73	34.92
Improvement(t1-t2)	13.46	13.07	13.27	13.14	13.20

Table 1 clearly shows the good performance of our initialization method. Our method uses less time to generate the same size of population and it can save more than 30% time compared to the traditional method.

It is shown in Table 2, the running time cost to generate 100000 individuals by employing traditional method (t1) and our new method (t2). We can find that the running time is slightly changed with the external degree of a node.

4.2 Experiments Results of NNIA-Net

In this section we study the effectiveness of our approach on a synthetic data set. Then we compare the results obtained by NNIA-Net with the Girvan and Newman's algorithm, in the following referred as GN, on some real-worlds networks for which the partitioning in communities is known. In both cases we show that our non-dominated neighbor immune algorithm successfully detects the network structure and is competitive with that of Girvan and Newman.

4.2.1 Evaluation Metrics

The evaluation metrics we use in this paper were also introduced in [3, 4].

An external measure, the Normalized Mutual Information(NMI) was adopted to estimate the similarity between the true partitions and the detected ones, and an internal one, the modularity introduced by Girvan and Newman. The Normalized Mutual Information is a similarity measure proved to be reliable by Danon et al. [1]. Given two partitions A and B of a network in communities, let C be the confusion matrix whose element C_{ij} is the number of nodes of community i of the partition A that are also in the community j of the partition B. The normalized mutual information $I(A,B)$ is defined as:

$$I(A,B)=\frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}C_{ij}\log(C_{ij}N/C_{i.}C_{.j})}{\sum_{i=1}^{c_A}C_{i.}\log(C_{i.}/N)+\sum_{j=1}^{c_B}C_{.j}\log(C_{.j}/N)} \qquad (9$$

)

where $c_A(c_B)$ is the number of groups in the partition A (B), $C_{i.}(C_{.j})$ is the sum of the elements of C in row i (column j), and N is the number of nodes. If $A=B$, $I(A,B)=1$. If A and B are completely different, $I(A,B)=0$. The larger value of NMI represents the greater similarity between A and B.

The modularity of Newman and Girvan [18] is a well known quality function to evaluate the goodness of a partition. Let k be the number of modules found inside a network, the modularity is defined as:

$$Q=\sum_{s=1}^{k}[\frac{l_s}{m}-(\frac{d_s}{2m})^2] \qquad (10)$$

where l_s is the total number of edges joining vertices inside the module s, and d_s is the sum of the degrees of the nodes of s.

4.2.2 Synthetic Data Set

The synthetic data set we use here was the benchmark introduced in 4.1.1. Nine different networks for values of external degree of a node ranging from 0 to 8 were generated, and the NMI is used to measure the similarity between the detected ones and the true partitions.

It is shown in Figure 2(a) that the NMI, averaged over the 20 runs, for values of the exponent r ranging from 1 to 1.5.

It is pointed out that, independently the value of r, NNIA-Net is able to detect exact community structure for the external degree of a node ranging from 0 to 4 (NMI=1).

With the increasing of the external degree of a node, each node has less links inside its community and more links with the rest of the network and it is more and more difficult to detect the structure within them (the values of NMI decrease).

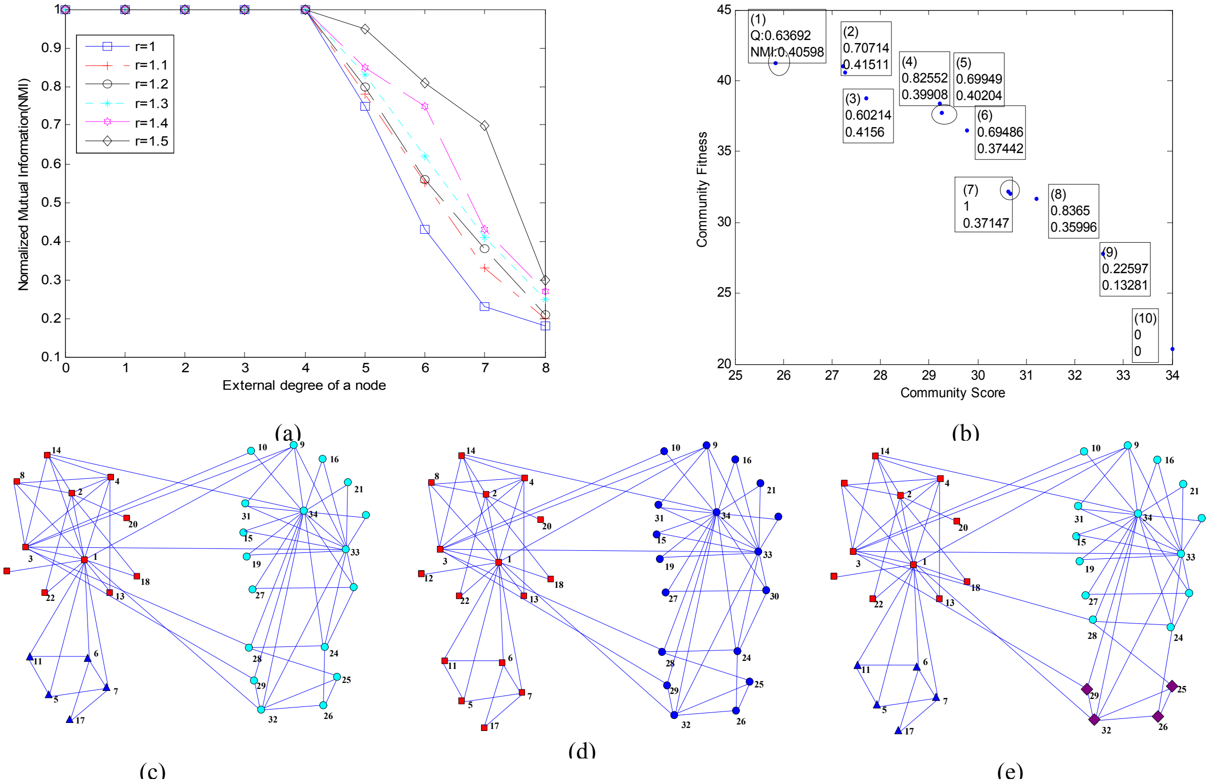
However, when the external degree of a node increases, higher values of r help in the retrieval of the true community structure.

4.2.3 Real-life Data Set

We now show the application of NNIA-Net on two real-world networks same as in [3, 4], the Zachary's Karate Club, the Bottlenose Dolphins, and compare our results with those obtained by Girvan and Newman's algorithm (GN).

Figure 2(b) displays the Pareto front in one out of the 10 runs, the network corresponding to the best value of NMI=1 with the modularity=0.37147(solution (7)), the one with the NMI=0.82552 and modularity=0.33908(solution (4)) and the one with the NMI=0.70714 and modularity=0.41511(solution (2)) . We can find that the solutions of the Pareto front have a hierarchical structure. Each of these solutions corresponds to a different trade off between the two objectives and thus to diverse partitioning of the network consisting of various number of clusters. The true partitioning which is displayed in Figure 2(d), consists of two modules obtained by the split of the two main groups. It is shown in Figure 2(c) that the left sub-graph is divided into two smaller ones and in Figure 2(e) that both the sub-graphs are divided into two smaller ones respectively (solution (2)).

The algorithm was executed 10 times on the two real-life data sets. At each run, the solutions having the best value of NMI and the best value of modularity have been selected. For each of them the corresponding modularity and NMI values, respectively, have been computed.



(a)

(b)

(c)

(d)

(e)

Figure 2. (a) NMI obtained by NNIA-Net on the synthetic network for different values of the exponent r.(b) Pareto front of one run. (c) Network corresponding to solution (4). (d) Network corresponding to the exact solution (node number (7) on the Pareto front). (e) Network corresponding to solution (2).

Table 3. Best NMI results obtained by our method and Girvan and Newman's algorithm for the real-life data sets

	avg best NMI	**std best NM**	**avg Mod**	**std Mod**	**GN NMI**
Zackary's Karate Club	1	0	0.371	0	0.692
Bottlenose Dolphins	1	0	0.373	0	0.573

Table 4. Best modularity results obtained by our method and Girvan and Newman's algorithm for the real-life data sets

	avg best Mod	std best Mod	avg NMI	std NMI	GN Mod
Zackary's Karate Club	0.415	0.02e-16	0.606	0.005e-15	0.380
Bottlenose Dolphins	0.505	0.017	0.528	0.018	0.495

The values are shown in Table 3 and Table 4. The tables clearly show the excellent performance of NNIA-Net with respect to Girvan and Newman's approach. At each run, the true partitioning to the two real-world data sets are obtained employing our algorithm NNIA-Net.

5. CONCLUSIONS

The paper presents a community detection algorithm based on NNIA, whose main idea is to optimize two objective functions which can evaluate a partitioning to a network. What's more, a new initialization method to generate safe individuals is firstly proposed in our paper. The experimental results show that our new initialization method is able to save as much as 30% time compared with the traditional initialization method and NNIA-Net has the capability to provide reasonable solutions to community detection problem.

6. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 60703107), the National High Technology Research and Development Program (863 Program) of China (Grant No. 2009AA12Z210), the Program for New Century Excellent Talents in University (Grant No. NCET-08-0811), the Program for New Scientific and Technological Star of Shaanxi Province (Grant No. 2010KJXX-03), and the Fundamental Research Funds for the Central Universities (Grant No. K50510020001).

7. REFERENCES

[1] Andrea Lancichinetti, Santo Fortunato, and Janos Kert'esz. 2008. *Detecting the overlapping and hierarchical community structure of complex networks.* arXiv: 0802.1281v1 [physics .soc-ph].

[2] Chuan Shi, Yi Wang, Bin Wu, and Cha Zhong. 2009. *A New Genetic Algorithm for Community Detection.* Complex 2009, Part II, LNICST, pp.1298-1309.

[3] Clara Pizzuti. 2008. GA-net:a genetic algorithm for community detection in social networks. In *Proc. of the 10th Intenational Conference on Parallel Problem Solving from Nature* (PPSN2008), pp.1081-1090.

[4] Clara Pizzuti. 2009. *A Multi-Objective Genetic Algorithm for Community Detection in Networks*. Tools with Artificial Intelligence, 2009. ICTAI '09. 21st International Conference, pp.379-386.

[5] Deb, K., Pratap, A. Agarwal,S.,and Meyarivan. T. 2002. *A fast and elitist multiobjective genetic algorithm: NSGA-II.* IEEE Transactions on Evolutionary Computation, 6(2): 182–197.

[6] E.A. Leicht and M.E.J. Newman. 2007. *Community structure in directed networks*. arXiv:0709.4500v1 [physics. data-an].

[7] Fortunato S and Barthélemy M. 2007. Resolution limit in community detection. In *Proc. Natl Acad Sci U S A*. 104(1): 36-41.

[8] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. 2004. Defining and identifying communities in networks.In *Proc. Natl .Acad.Sci. USA*(PNAS'04),101(9): 2658-2663.

[9] Julia Handle and Joshua Knowles. 2007. *An evolutionary approach to multiobjective clustering*. IEEE trans-actions on Evolutionary Computation, 11(1): 56–76.

[10] Kumpula JM, Saramaki J, Kaski K and Kertesz J. 2007. *Resolution limit in complex network community detection with Potts model approach.* The European Physical Journal B, Condensed Matter and Complex Systems, pp.41-45.

[11] M.E.J. Newman. 2004. *Detecting community structure I in networks*. Eur.Phys.J.B, 38(2):321.

[12] M. Girvan and M.E.J. Newman. 2002. Community structure in social and biological networks. In *Proc.National.Academy of Science.* USA99, pages7821-7826.

[13] M. Girvan and M.E.J. Newman. 2002. Community structure in social and biological networks. In *Proc.National.Academy of Science.* USA99, pages7821-7826.

[14] Mursel Tasgin, Amac Herdagdelen, and Aluk Bingol. 2007. *Communities detection in complex networks using genetic algorithms*.oai:arXiv.org:0711.0491v1[physics.socph],2007.

[15] Maoguo Gong, Licheng Jiao, Haifeng Du, Liefeng Bo. 2008. *Multiobjective Immune Algorithm with Nondominated Neighbor-Based Selection*. EC by the Massachusetts Institute of Technology, Summer 2008, Vol. 16, No. 2: 225-255.

[16] Rumi Ghosh, Kristina Lerman. 2008. *Community Detection using a Measure of Global Influence*. The 2nd SNA-KDD Workshop'08.

[17] S.Lozano, J.Duch, and A.Arenas. 2007. *Analysis of large social data sets by community detection*. European Physical Journal ST(143): 257-259.

[18] Wasserman S. Faust K. 1994. *Social Network Analysis: Methods and Applications.* Cambridge University Pre.

[19] Y.J.Park and M.S.Song. 1989. A genetic algorithm for clustering problems. In *Proc.of 3rd Annual Conference on Genetic Algorithms*, pp.2–9.