PSO-based Community Detection in Complex Networks

Zhewen Shi[1], Yu Liu[1], Jingjing Liang[2]
1 School of Software, Dalian University of Technology, Dalian, China
2 School of Computer and Information Technology, Nanyang Normal University, Nanyang, China
e-mail: zhewen.shi@gmail.com

***Abstract*—Community detection is always an outstanding problem in the study of networked systems such as social networks and computer networks. In this paper, a novel method based on particle swarm optimization is proposed to detect community structures by optimizing network modularity. At the beginning, an improved spectral method is used to transform community detection into a cluster problem and the weighted distance which combine eigenvalues and eigenvectors is advanced to measure the dissimilarity of two nodes. Then, PSO is employed for cluster analysis. There are two definitive features in our algorithm: first, the number of communities can be determined automatically; second, the particle has low-dimensional structure by using only the corresponding components of the first nontrivial eigenvector to express community centers. The application in three real-world networks demonstrates that the algorithm obtains higher modularity over other methods (e.g., the Girvan-Newman algorithm and the Newman-fast algorithm) and achieves good partition results.**

Keywords-community detection; spectral method; particle swarm optimization; modularity

1. INTRODUCTION

Many actual issues can be represented as networks, such as social networks, computer networks, neural networks and communication networks. The research [1] by Girvan and Newman showed that community structure was the common property of many networks and community detection was defined as dividing network nodes into groups within which there were dense network connections, but between which the sparser ones. Especially with the development of Internet, community detection in complex networks, as the basis of personal web service, becomes more and more important.

Community detection has attracted much attention and several algorithms have been proposed, such as Girvan-Newman (GN) algorithm [1], Newman-fast algorithm [2] and the DA [3]. In recent years, with the widely application of computational intelligence (CI), some algorithms based on CI have been utilized in detecting community structure. In [4] and [5], genetic algorithm (GA) and particle swarm optimization (PSO) were used separately. However, the cluster each vertex belongs to was encoded into one individual, which made the dimension high and cost much time during the process. Liu [6] adopted a spectral method to transform the issue into a cluster problem and then used GA to detect community structure. Although only cluster centers were encoded into each chromosome, no relation between nontrivial eigenvectors obtained from the spectral method was considered and the number of communities was obtained by exhaustive search, which also cost much time.

In this paper, Capocci algorithm, an improved spectral method, is utilized first to transform community detection into a cluster issue and the weighted distance is proposed to measure the dissimilarity of two nodes. Then, PSO is used to optimize the network modularity. During the process, the number of clusters can be obtained automatically.

The rest of this paper is organized as follows. In Section2, the spectral method weighted distance is proposed, and its effect of measuring the dissimilarity of different nodes is also discussed. In Section 3, the whole process of PSO-based community detection algorithm is described in detail, especially the structure of a particle. In Section 4, experimental results in three real world networks are shown and compared with the performance of other classical methods. Finally, Section 5 gives conclusions.

2. SPECTRAL METHOD WEIGHTED DISTANCE

2.1 Capocci Algorithm

Based on the analysis of Normal matrix, Capocci developed an improved spectral method [7] for community detection. The Normal matrix is defined as $N = D^{-1}W$, where W is the adjacency matrix, D is the diagonal matrix with element $D_{ii} = \sum_{j=1}^{s} W_{ij}$ and s is the number of vertices in the network. Capocci pointed out that the matrix N always had the largest eigenvalue equal to 1, associated to a trivial constant eigenvector. If a network has an apparent cluster structure and consists of m communities, N has m-1 nontrivial eigenvalues $A_p (p = 1, \cdots, m-1)$ close to one, the other nontrivial eigenvalues lie a gap away from 1. In the first m-1 nontrivial eigenvectors $V_p = (v_{p1}, v_{p2} \cdots, v_{ps})$ $(p = 1, \cdots, m-1)$ associated to A_p, the components corresponding to nodes within the same community have similar values. As a result, communities can be detected by each eigenvector $V_p (p = 1, \cdots, m-1)$ if the network structure is clear enough.

2.2 Spectral Method Weighted Distance

In most common networks, no clear partition exists. In other words, the profile of the each eigenvector V_p, sorted by components, is smooth and not step-like. For example, Fig. 1 shows Zachary Karate Club network [8], which consists of 34 vertices and 78 edges. Fig. 2 shows the components of the first nontrivial eigenvector V_1. It appears that although the value of 0 can correctly partition the network into two communities, the eigenvector profile is too smooth for clear partition.

978-0-7695-3888-4/09 $26.00 © 2009 IEEE
DOI 10.1109/KAM.2009.195

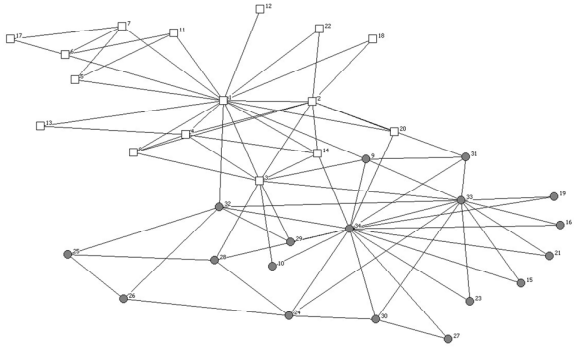

Figure 1. Zachary Karate Club Network.

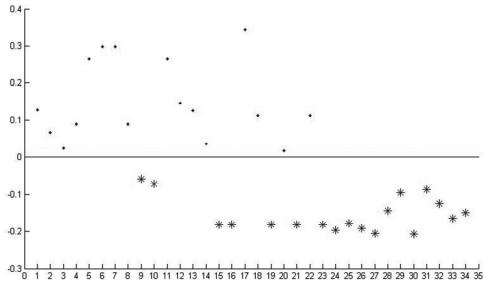

Figure 2. Zachary Karate Club Network: values of the first nontrivial eigenvector V_1 components.

Take American College Football Network [1] for example, it consists of 115 vertices and 616 edges, the vertices in the network are the college football teams in 12 communities and the edges are matches between two different teams. Fig. 3 shows the distribution of the first 11 nontrivial eigenvectors $V_p(p=1,\cdots,11)$. It appears that neither the number nor member teams of communities can be found correctly from each eigenvector $V_p(p=1,\cdots,11)$. Fig. 4 shows components of the first nontrivial eigenvector V_1 corresponding to nodes in Big Twelve and Mountain West conference. It is clear that components corresponding to the same conference's teams are not close enough in one eigenvector if the network structure is not clear.

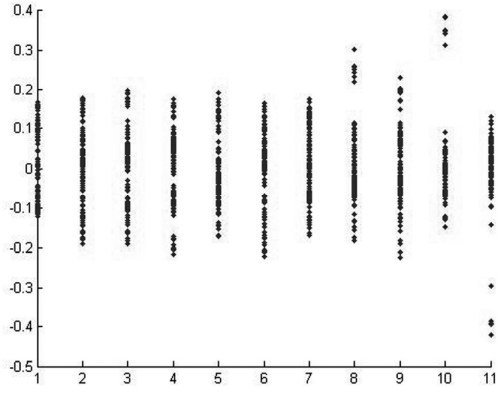

Figure 3. American College Football Network: values of the first 11 nontrivial eigenvectors components.

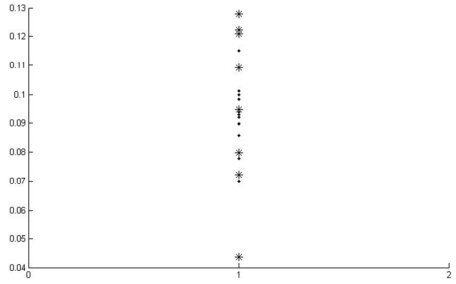

Figure 4. American College Football Network: components of the first nontrivial eigenvector V_1 corresponding to nodes in two conferences (*teams in Mountain West, • teams in Big Twelve).

Since there is no clear partition in most networks and none of the nontrivial eigenvectors can be used alone to divide vertices, a new weighted distance which combine eigenvalues and eigenvectors is defined as follows:

$$Dis_{i,j} = \sqrt{\sum_{k=1}^{m-1} A_k \times (v_{ki} - v_{kj})^2} \qquad (1)$$

Where $Dis_{i,j}$ measures the dissimilarity of i-th node and j-th node and a small $Dis_{i,j}$ means there is a big possibility that the i-th node and j-th node belong to the same community.

Take the twenty teams in Big Twelve conference and Mountain West conference for example, Fig. 5 shows their dissimilarities obtained by (1), in which m is set to 12. 1 to 12 in Fig. 5 represent the twelve teams in Big Twelve conference and 13 to 20 represent the eight teams in Mountain West conference. It appears that the weighted distance between teams in the same community is much smaller than that between teams in different communities.

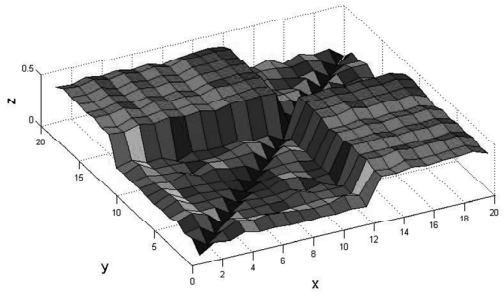

Figure 5. The dissimilarity between teams in Big Twelve and Mountain West.

In addition, eigenvalues in (1) enhance the robustness of spectral method weighted distance. Since the smaller eigenvalue reduces the proportion of corresponding eigenvector to the whole distance, little change about the max of k will not affect the value of weighted distance much.

3. PSO-BASED COMMUNITY DETECTION

Since each network has its own number of communities, the value of m in (1) can hardly be obtained by professional experience. However, it is much costly to

obtain m by exhaustion algorithm. In this paper, suppose (r+2)-th eigenvalue is the first negative one, the first r nontrivial eigenvectors obtained by Capocci Algorithm are selected for community detection. Because r is always no less than m-1, r+1 is set to the max number of communities and the appropriate value of r is obtained during the process of community detection.

The matrix T made up by the first r nontrivial eigenvectors is defined as follows:

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_s \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1r} \\ t_{11} & t_{12} & \cdots & t_{1r} \\ \vdots & \vdots & \vdots & \vdots \\ t_{s1} & t_{s2} & \cdots & t_{sr} \end{bmatrix} \quad (2)$$

Where each column represents one nontrivial eigenvector and each row represents one node in the network.

3.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO), a new evolutionary computation technique inspired by social behavior simulation of bird flocking, was introduced by Kennedy and Eberhart in 1995[9]. Because of its fast convergence and promising performance, PSO has attracted much attention and been widely used in complex function optimization, neural network training and data mining. Among variations of PSO, LDWPSO (the linear weighted PSO) [10] is selected in this paper.

A swarm consists of m particles moving around in a D-dimensional search space $[X_{\min}, X_{\max}]^D$. The i-th particle at the t-th iteration has a position $X_i^{(t)} = (x_{i1}, x_{i2}, \cdots x_{iD})$, a velocity $V_i^{(t)} = (v_{i1}, v_{i2}, \cdots, v_{iD})$, the best solution achieved so far by itself (*pbest*) $P_i = (p_{i1}, p_{i2}, \cdots, p_{iD})$. The best solution achieved so far by the whole swarm (*gbest*) is represented by $P_g = (p_{i1}, p_{i2}, \cdots, p_{gD})$. The position of the i-th particle at the next iteration will be calculated to the following equations:

$$v_{id}^{(t+1)} = w \cdot v_{id}^{(t)} + c_1 r_1 (p_{id} - x_{id}^{(t)}) + c_2 r_2 (p_{gd} - x_{id}^{(t)}) \quad (3)$$

$$x_{id}^{(t+1)} = x_{id}^{(t)} + v_{id}^{(t+1)} \quad (4)$$

Where c_1 and c_2 are two positive constants; r_1 and r_2 are two random numbers in the range [0,1]; w is inertia factor which linearly decreases from 0.9 to 0.4 through the search process. In addition, the velocities of the particles are confined within $[V_{\min}, V_{\max}]^D$, where $V_{\min}$ is always equal to $X_{\min}$ and $V_{\max}$ is always equal to $X_{\max}$. If an element of velocities exceeds the threshold $V_{\min}$ or $V_{\max}$, it is set equal to the corresponding threshold.

3.2 Encoding Particle Structure

In order to determine the structure of a network, the vertices should be clustered and each cluster is a community. The choice of an efficient representation for network structure is one of the most important issues in the optimizing process. If all vertices are encoded into a particle, the length of the particle is too long for quick convergence. In addition, complexly operation should be done at the end of the algorithm to eliminate the alone vertices. Therefore, only the number of communities and the center of each community are encoded. The structure of the particle includes two parts are as follows:

Center existence array				Center array			
$flag_1$	$flag_2$	··· ···	$flag_{r+1}$	$center_1$	$center_2$	··· ···	$center_{r+1}$

Figure 6. Structure of a particle.

Where $flag_i$ indicates whether or not the i-th center is valid in the result, their values are between 0 and 1. If $flag_i < 0.5$, the i-th center is not included in the result partition. Otherwise the i-th center is valid. In such a way, the number of valid $flag_i$ represents the number of communities.

$center_i$ is the i-th community center corresponding to $flag_i$. If $flag_i$ is valid, $center_i$ is valid too. Otherwise, $center_i$ is not used when computing partition result. Since the components of t_i in (2) are interrelated and t_{i1} represents the feature of i-th vertex best in t_{ij} ($j = 1, 2, \cdots, r$), only the first dimension of the cluster center is encoded. Suppose $a = \min(t_{11}, t_{21}, \cdots, t_{s1})$ and $b = \max(t_{11}, t_{21}, \cdots, t_{s1})$, then $center_i \in [a, b]$.

Each particle in the swarm expresses a partition of vertices in the network. When transforming a particle into a factual network partition, the j-th node within which t_{j1} is closest to the valid $center_i$ among all t_{p1} ($p = 1, 2, \cdots, s$) is a valid center of communities. After all centers are determined, the spectral method weighted distances from nodes to all centers are computed and each node is divided to the community with the smallest weighted distance. The process is described as follows:

1) Compute the number of valid $flag_i$, *let* $n = \sum_{i=1}^{r+1} round(flag_i)$.

2) If n<2, n=2, find p and q where $flag_p$ *and* $flag_q$ *are no less than other* $flag_i (i \in [1, r+1])$ *,make* $flag_p = 1 - 0.5 \times rand(0,1)$, $flag_p = 1 - 0.5 \times rand(0,1)$.

3) For each $flag_i > 0.5$, *find* $t_{j1} \in T$ *where* t_{j1} *is closest to* $center_i$. *Make* $[t_{j1}, t_{j2}, \cdots, t_{j(n-1)}]$ *one valid center of the communities.*

4) Combine centers to one if they have the same value. If the modified n=1, then set n=2 and select $t_{j'1}$ *where* $t_{j'1}$ *is furthest from the original center* t_{j1}. *Replace the combined* $center_i$ *with* $t_{j'1}$ *and set* $[t_{j'1}, t_{j'2}, \cdots, t_{j'(n-1)}]$ *the second center.*

5) Compute spectral method weighted distance from all vertices to centers by (1),where m is set to n. Divide nodes to communities with the smallest distance to their centers.

The variation in genetic algorithms is advanced in step *2)* and step *4)*. When the number of communities obtained from a particle is smaller than 2, some particles are mutated to maintain the number no less than 2.

3.3 Designing fitness function

The fitness function guides the evolution process. In this paper, we use network modularity, a quantitative definition proposed by Girvan and Newman [11] to measure the performance of the partition. A big fitness denotes a good particle.

Suppose the number of communities obtained from a particle is n, define a $n \times n$ symmetric matrix $E = (e_{ij})$, where e_{ij} indicates the fraction of all edges in the network that link vertices in community i to the vertices in community j. Then a $1 \times n$ matrix $A = [a_1, a_2, \cdots, a_n]$ is defined, where $a_i = \sum_{j=1}^{n} e_{ij}$ indicates the fraction of edges connect to vertices in community i. The modularity is described as follows [11]:

$$Q = \sum_{i=1}^{n} (e_{ii} - a_i^2) \qquad (5)$$

Where the value of Q is between 0 and 1, higher value represents better partition.

3.4 Optimizing Process

The pseudocode of the PSO-based process is defined in Fig. 7.

```
Initialize all particles in the swarm and their velocities;
For t=1 to the limit of iterations
{
    For each particle i in the swarm
    {
        Transform the particle to a network partition;
        Compute its fitness according to (5);
        Update its velocity according to (3);
        Update its position according to (4);
        If possible, update P_i and P_g;
    }
    Terminate if P_g meets problem requirements;
}
```

Figure 7. The pseudocode of the optimizing process.

4. EXPERIMENTS

Here we evaluated our algorithm on three well known datasets with known community structures. Our aim was to prove the structure of PSO-based community detection validity, so only the partition results were listed and the parameters of PSO were not fine tuned. Some common parameters were set as follows: the swarm size 20, the limit of iterations 1000, weight w decreasing linearly between 0.9 and 0.4, learning rate $c_1 = c_2 = 2.05$. For each experiment, 30 runs of the algorithm were performed.

4.1 Karate Network

Karate Network consists of 34 vertices and 78 edges. The first three nontrivial eigenvalues of Normal matrix, which were 0.8672, 0.7044, 0.6130 respectively, were found positive. So the max number of communities was set to 4.

According to the structure of a particle, the swarm moved around in 8-dimensional search space, where the first four dimensional space were between 0 and 1, the last four dimensional space [-0.206, 0.3431] were between the min and max components of the first nontrivial eigenvector. In all of the runs, the results partition (Fig. 8) consisted of 4 communities and the modularity was 0.4199.

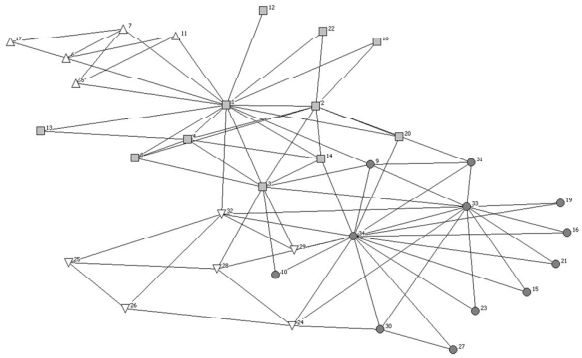

Figure 8. Partition result of Karate Network by our algorithm.

Compared with Fig. 1, each cluster was divided into two and a higher modularity 0.4199>0.3715 was obtained. Table 1 listed the number of communities and the maximum modularity achieved by our algorithm compared to the results listed in GN [1], Newman-fast [2], DA [3] and spectral algorithm [13]. It can be seen clearly that our algorithm got the highest modularity.

TABLE I. THE PARTITION RESULTS OF KARATE NETWORK

Algorithms	Number of communities	Modularity
GN	3	0.401
Newman-fast	2	0.381
DA	4	0.4188
Spectral algorithm	4	0.4188
Our algorithm	4	0.4198

In addition, as described in original papers, the node 10 was classified incorrectly in Newman-fast, DA and spectral algorithm, node 3 was misplaced in GN. However, all vertices were classified correctly in our algorithm.

4.2 Dolphins Network

The Dolphins Network (Fig. 9), which consists of 62 nodes and 159 edges, was complied by Lusseau [12] from seven years studies of dolphins living in Doubtful Sound. Each node is a dolphin and an edge indicates frequent association between dolphins.

In the corresponding normal matrix, the first four nontrivial eigenvalues were positive. So the first four nontrivial eigenvectors were selected and the max number of communities was 5.

In all runs, the modularity was between 0.516 and 0.5249. The maximum modularity was 0.5235 when splitting into four groups and 0.5249 when splitting into five groups. Fig. 10 showed the partition result with the modularity 0.5249. Compared with Newman's result

0.52 ± 0.03 in [11], our algorithm got the higher modularity. Additionally, the two algorithms got the same result when combining the three groups on the right side of Fig. 10 into one.

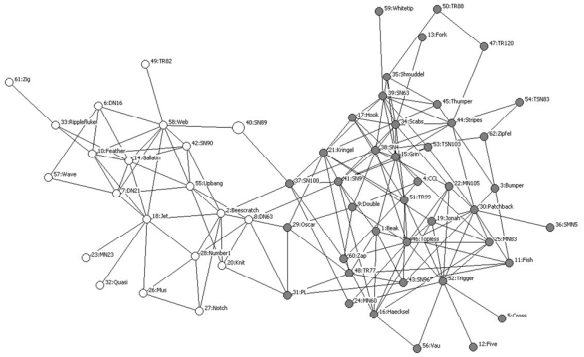

Figure 9. Dolphins Network.

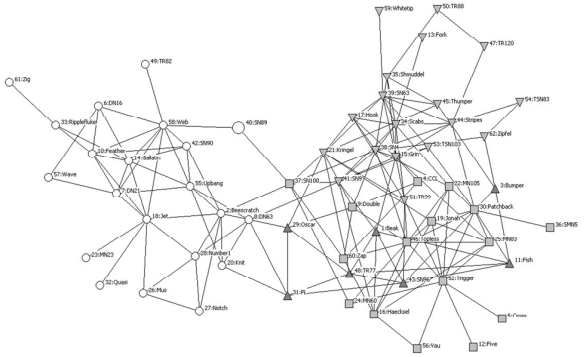

Figure 10. Partition result of Dolphins Network by our algorithm.

4.3 American College Football Network

After Capocci algorithm was executed, the first 14 nontrivial eigenvectors were selected for community detection. In all the runs, the results partition consisted of 10 communities and the modularity was 0.6046.

From Fig. 11, we can see that only the Sunbelt conference and IA Independents conference were detected incorrectly. Table 2 listed the number of communities and the maximum modularity achieved by our algorithm compared to the results listed in GN [1] and Newman-fast [2]. It is clear that our algorithm got the higher modularity.

TABLE II. THE PARTITION RESULTS OF AMERICAN COLLEGE FOOTBALL NETWORK

Algorithms	Number of communities	Modularity
GN	12	0.601
Newman-fast	6	0.546
Our algorithm	10	0.6046

Compared with the partition result represented in Newman-fast, our algorithm separated Atlantic Coast from Big East, Conference USA from SEC, Mountain West from Pacific10, and got a higher accuracy in Mountain West conference. Compared with the partition result represented in GN, our algorithm only grouped the cluster of node 11, 24, 50, 69 into Mountain West conference, and grouped the cluster of node 59,63,97,58 into SEC conference. The accuracy of the partition was not influenced.

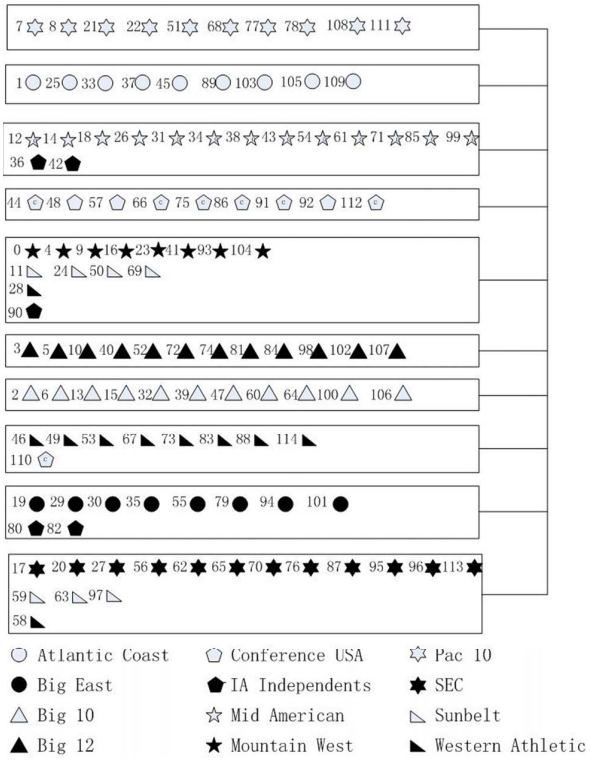

Figure 11. Partition result of American College Football Network by our algorithm (The ID number is as same as the number provided by Newman[14]).

5. CONCLUSIONS

In this paper, we proposed a new method based on particle swarm optimization for community detection. There are three innovations in this paper. Firstly, based on Capocci algorithm, spectral method weighted distance which combined eigenvalues and eigenvectors is advanced to measure the dissimilarity of two vertices. Since smaller eigenvalues make less contribution to the distance, this measure enhanced the robustness when the exact number of communities can hardly be found. Secondly, the dimension of particles is compress by using components of the first nontrivial eigenvector to describe the community centers. Finally, the number of communities is encoded impliedly in the particle. It does not need any prior knowledge and the number of communities can be determined automatically. Experimental results show that our algorithm achieves better performance than some other methods (e.g., the Girvan-Newman algorithm and the Newman-fast algorithm).

ACKNOWLEDGMENT

It is a project supported by the Natural Science Foundation of China (No.60803074, 60673024) and the Natural Science Foundation of Liaoning Province (No.20082172).

REFERENCES

[1] M. Girvan, M. Newman. Community structure in social and biological networks. Proceeding of the National Academy of Sciences, 99(12):7821-7826, 2002.

[2] J. ME. Fast algorithm for detectiong community structure in networks. Physical Review E, 69:066133, 2004.

[3] J. Duch, A. Arenas. Community detection in complex networks using extremal optimization. Physical Review E, 72(2):27104, 2005.

[4] X. Liu, D. Li, S. Wang, and Z. Tao. Effective algorithm for detecting community structure in complex networks based on GA and Clustering. Lecture Notes in Computer Science, 4488:657, 2007.

[5] X. Duan, C. Wang, X. Liu, Y. Lin. Web community detection model using particle swarm optimization. Computer Science, 35(003):18-21, 2008.

[6] T. LIU, B. HU. Detecting community in complex networks using cluster analysis. Complex Systems and Complexity Science, 4(1):28-35, 2007.

[7] A. Capocci, V. Servedio, G. Caldarelli, and F. Colaiori. Detecting communities in large networks. Physica A: Statistical Mechanics and its Applications, 352(24):669-676, 2005.

[8] W. Zachary. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 452-473, 1977.

[9] J. Kennedy, R. Eberhart. Particle swarm optimization. In IEEE International Conference on Neural Networks, 1995. Proceedings, volume 4, 1995.

[10] Y. Shi and R. Eberhart. A modified particle swarm optimizer. In Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, 69-73, 1998.

[11] M. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):26113, 2004.

[12] D. Lusseau and M. Newman. Identifying the role that animals play in their social networks. In Proc. R. Soc. Lond. B (Suppl.), 271: 477-481, 2004.

[13] M. Newman. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.

[14] http://www-personal.umich.edu/ mejn/netdata/.