Computer Science and Information Systems 12(2):843–855 DOI: 10.2298/CSIS141021029C

Detecting Overlapping Community in Complex Network Based on Node Similarity

Zuo Chen[1,2], Mengyuan Jia[1], Bing Yang[3], and Xiaodong Li[1]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, China
chenzuo@iie.ac.cn
[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[3] School of Education, Hubei University, Wuhan, Hubei 430062, China
yangbing@126.com

Abstract. Overlapping communities in complex network is a common phenomenon in real world network. The overlapping community structure can more accurately obtain the actual structure information in the network. But at present the study of overlapping community division algorithm is relatively less, facing the problems of the low accurate rate. Based on this, this paper presents algorithms OCNS for detecting community overlapping base on node similarity. The algorithm calculates similarity between two nodes in the network by means of Jaccard similarity measure formula. Then the related nodes are adaptive merged according to the similarity value, combining with the community according to the change of modularity. The process of partitioning can not only accurately merge closely linked nodes in the network, but also find the overlapping nodes and bridge nodes between communities. The experiment proved the algorithm is effective to detect the overlapping community and has obvious advantages in the division of baseline social network Zachary and dolphin network, and the quality of division better than other existing partitioning algorithm.

Keywords: complex network, community structure, node similarity, modularity.

1. Introduction

Nature is a complex, common relationship and interaction system. The common feature of these networks is a complex internal structure. If it regards the individual as nodes, relations among individuals as edges, then we can get complex network which has a large number of nodes and intricate relationship. In real life, many complex systems can be represented by complex network, such as social network, Internet etc. As one of the basic characteristics of complex networks, community structure exists in various complex networks generally, also one of most basic research in this filed. Community structure [1] is that the internal of network contacts closely and connects with the other part of the network is loose. In complex networks, the community has a certain function and real existence. For example, communities in social networks can represent

an interest group or an organization. In Webpage web may represent a topic. Therefore, it is a very important work to detect and analyze community structure accurately.

On the whole, Community detecting algorithm is divided into non overlapping partitioning algorithm and the overlapping community division algorithm. Non overlapping algorithm is mainly based on graph partitioning and hierarchical clustering. Among them, the famous graph partitioning algorithm are the Kernighan-Lin algorithm [2] and the Laplace [3] chart eigenvalue based on spectral score. But these two algorithms can only divides the network into two community of known size. The community of complex network of real world are often vague and need to artificial mining.

Hierarchical clustering algorithm achieve the natural segmentation of community structure by adding the network sides or cutting edges to. Therefore, it includes agglomerative method and splitting method. In 2002, Newman and Girvan proposed GN algorithm [4] and the modularity function to evaluate the goodness of partition. Then some detecting community methods have been proposed based on the optimizing modularity [5-7]. Aggregation algorithm mainly include CNM algorithm [8] proposed by Newman in 2004. Ref. [9] proposed BGLL algorithm etc. But these algorithms are detecting non overlapping algorithms. Considering the actual network, there often exists overlapping community. While both the graph partitioning algorithm, split or agglomerative algorithm, are ultimately divided the network into several independent community structure without considering the existence of overlapping community. The distribution is contrary to the actual network. And some methods still lack the optimal termination conditions , leading to decreasing the accuracy.

Research on the overlapping community detecting is less relative to the non overlapping community. CMP algorithm [10] presented for the analysis of overlapping community. An improved CONGA algorithm [11] presented based on GN detected the overlapping community. Ref. [12] can find the overlapping community and hierarchical structure based on local modularity optimization. The algorithm is one of the representative algorithm to divide community structure based on the theory of extremely optimization. Ref. [13] detect overlapping communities based on the connection degree, whose results is more accurate. But there still exists many problems, such as the efficiency problem of the algorithm, the algorithm parameter selection problem etc..

In view of research status of community detection at present, we put forward a method OCNS to detect the overlapping community based on node similarity. The algorithm merges the related nodes with larger similarity into a community and select the sub community which have the maximum number of nodes as a baseline community and finally adaptively optimize the modularity to get the best partitioning community. It has a linear time complexity, and compared with other algorithms, OCNS is more effective and can get more accurate overlapping nodes.

The rest of this paper is organized as follows: First the concept about community are briefly introduced in Section 2; Then the proposed method OCNS is described detailed in Section 3 and the simulation results are presented in Section 4. Finally, the conclusions and comments are drawn in Section 5.

2. The Related Concepts

Before describing the algorithm OCNS, we introduce some related knowledge of community structure. Complex network community detection is a challenging problem. The network of detecting are mostly simple graph. Suppose there is a network G with N nodes and M directed edges. Let A be the adjacency matrix of the network G with N nodes, as shown in formula (1):

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} \quad (1)$$

where the element A_{ij}=1 denotes that there is an edges from node i to node j, otherwise A_{ij}=0. Elements of A can be defined as follows:

$$A_{ij} = \begin{cases} 1, & if \ node\ i\ and\ j\ are\ connected \\ 0, & otherwise \end{cases} \quad (2)$$

2.1. Node Degree and Neighbor Node

In network G, the degree of node i is a number of edges associated with node i. The research on the degree distribution of nodes is important for the assessment of the network. It generally believes that the greater the degree of node, node is more important. For undirected network, without considering node's in-degree and out-degree. The degree of node i with community C is defined as the number of edges of node i and the nodes of community C. Intuitively, the greater the degree, node i and community C is more closely.

The neighbor of the node i in network is the set of directly connected with the node i, then the neighbor node set of community C can be defined as:

$$Neibor_C = \bigcup_{i=1}^{n} Neibor_i \quad (3)$$

The first step of this algorithm is to search the neighbor nodes of all nodes.

2.2. Overlapping Nodes and Bridge Node

If a node simultaneously in two or more than two communities, then the node is a overlapping node between the community.

For any node, if the node is connected with two or more communities, and only one side, the node is considered bridge node. As can be seen from the definition, the bridge node occupies an important strategic position in network communication behavior. At the same time, the bridge node must be overlapping nodes, in turn, overlapping nodes do

not always bridge node. As shown in figure 1, Black and yellow represent the two communities, node 6 and node 8 is the common node of two communities. However, node 8 is the bridge node, node 6 is not.

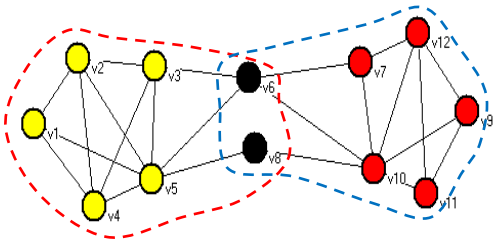

Fig. 1. Overlapping nodes and bridge node. Two communities labeled by yellow and red circle, overlapping part labeled by black circle, where node v_8 is also a bridge node

2.3. Modularity Function *Q*

Community modularity Q is a parameter which used to characterize the strength of the association, and it is the most widely used index of evaluating the strength of the community characteristics. It is defined as follows:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j) \qquad (4)$$

where k_i and k_j is the degree of node i and node j, C_i is the affiliated community of node i, m represents the total number of edges of the network. When $C_i=C_j$, $\delta(C_i,C_j)=1$, otherwise equals 0. The Q value is between 0~1 and the real network modularity function value is generally between 0.3~0.7. In generally, making Q=0.3 is the lower bound of having obvious community in network. The greater the value of the modularity, community structure is more obvious.

3. The Proposed Method OCNS

3.1. Node Similarity

Classification or clustering often needs to estimate the similarity between different samples, and the usual method is calculating the distance between samples. It is very important to adopt appropriate methods and related to the classification correctness. The commonly used "distance" method is Euclidean distance and correlation coefficient. The paper use the Jaccard correlation coefficient [14] to measure the close degree of each

pair of nodes, which use the normative public neighborhood size to measure the similarity between two nodes. Similarity is defined as follows:

$$Jaccard(i, j) = \frac{\Gamma(i) \bigcap \Gamma(j)}{\Gamma(i) \bigcup \Gamma(j)} \quad (5)$$

In network G, $\Gamma(i) \bigcap \Gamma(j)$ shows intersection of the neighbor nodes between node i and node j. $\Gamma(i) \bigcup \Gamma(j)$ shows union of the neighbor nodes between node i and node j. The ratio of them is the similarity between the two nodes. If the intersection of neighbor nodes of the two nodes is 0, then the similarity between the two nodes is 0. These nodes are often exist in the network.

For a network G of N nodes, by means of calculating the Jaccard similarity between each pairs of nodes, getting the similarity matrix S of network G. S is obviously a weight matrix. S is defined as follows:

$$S = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{21} & S_{22} & \dots & S_{2n} \\ \dots & \dots & \dots & \dots \\ S_{n1} & S_{n2} & \dots & S_{nn} \end{bmatrix} \quad (6)$$

In this paper, the relationship between the edges in the original network converted into the similarity value between nodes. Detailed algorithm process is as follows.

3.2. Algorithm Description

For indirection and un-weighted networks, a node can be similar to multiple nodes based on the relationship of network. Based on the concept of similar network, the two nodes should be divided into together if their similarity are enough big. We merge these nodes with larger similarity as the initial community and carry through directed merging according to the change of modularity. In the process of merging, The key lies in the analysis of the relationship between the nodes. If the modularity Q is larger when merging them, the two little communities are closely linked, and should be divided into together.

The method OCNS is based on the idea the some similarity nodes should have more chance to belong to the same community and some nodes should not only in a community. Then, the procedure of our method is as follows:

Step 1. Finding the neighbor nodes of all nodes from the network *G (V, E)* and calculating the similarity of each pairs nodes according to the Jaccard similarity formula, to obtain the similarity matrix S. Find out nodes whose the similarity value equals to 0 and put in set S_0.

Step 2. Based on the similarity of nodes, finding the most similarity of each node. Then combined with the relevant similarity nodes.

Step 3. The merged community number is set to *C1, C2, C3,...*, and calculate each community modularity *Q*. From the small community to select a community whose have the largest number of nodes as the initial community. Based on the community to merge

the rest small community, towards the direction of increasing of modularity value to merge.

Step 4. Then from the remaining community to find out a community which have the largest number of nodes. Repeat *step 3* until all of small communities complete the division. At this time, nodes with overlapping tendency will also divided into different communities.

Step 5. Calculating the degree between the nodes in the set S_0 and each community. Add them to the community whose have the maximum degree. Otherwise, join with a larger neighbor nodes. Finally, calculating the degree between the overlapping nodes and communities connected to find the bridge node.

Here, community detection is completed. To make our method clear to readers, we show specific algorithm flowchart as shown in figure 2 and gives the network in figure 1 as an example to explain our algorithm. Table 1 gives the most similar node of each node in figure 1 according to its similarity matrix.

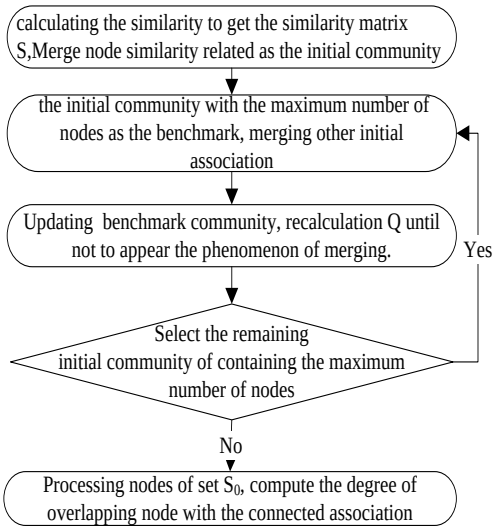

Fig. 2. Flow chart of algorithm

Table 1. Nodes and their neighbors with maximum similarity of the example network

Node	S_{max}	Node	S_{max}	Node	S_{max}
1	2	5	2	9	11
2	4	6	7	10	12
3	5	7	10	11	9
4	2	8	no	12	10

According to the above node information, combined with related the node which have larger similarity, as shown in figure 3:

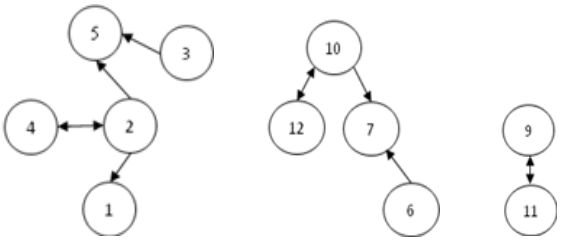

Fig. 3. Similar nodes merging

By similarity algorithm to get the size of the similarity of each node, The direction of the node similarity has been marked by arrows. The small network divide into 3 small associations. The network of 12 nodes has been reduced to 3. The algorithm OCNS is with these three initial community as initial associations to merge and detect the actual communities of the network. But when merging these related similar nodes, we must be moderate. Otherwise it will affect the efficiency and accurate of algorithm.

4. Experiment

In this paper, experimental platform is *matlab*R2012a. To evaluate the performance of the proposed method, some networks such as the Zachary's karate club network, dolphin association network and the computer-generated networks are used to be the test networks and compare with six algorithms which is GN, FN, BGLL, Lfm, algorithm proposed by Ref. [15] and Ref. [16] respectively.

4.1. Zachary's karate club network

The Zachary's karate club network [17] is one of the classic studies in social network analysis and has been used as one of the typical test examples by many researchers to detect community structures in complex network [12,18,19]. The club network consists of 34 member nodes, and splits in two smaller clubs after a dispute arose during the course of Zachary's study.

According to this algorithm, first step should search the neighbor nodes and get the similarity. Relevant similarity nodes of each node is as shown in figure 4 and detection results of the Zachary network is as shown in figure 5:

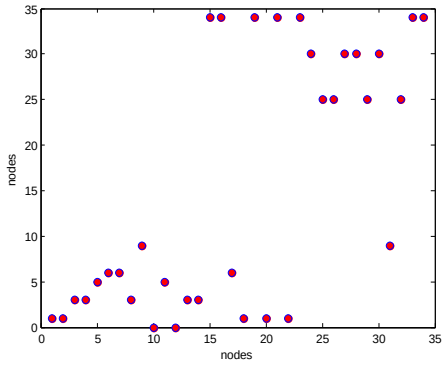

Fig. 4. Relevant similarity nodes of each node (Zachary)

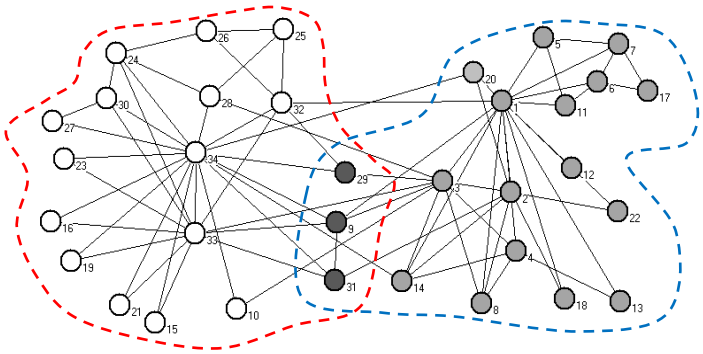

Fig. 5. Community structure of Zachary network. Two communities are detected, labeled by white and grey circle, and overlapping part labeled by black circle.

We can see from figure 4, node 34 and node 1 are similar to a part of nodes. So these relevant similarity nodes can be classified as small communities. The main part of the algorithm is also based on the initial small communities. Similarity value of node 10 and node 12 is equal to 0 and we should decide which community they are in the last. As mentioned in many literatures, division of node 10 is controversial, we can also prove this point from the similarity curve. In our method OCNS, due to node 10 has only two neighbors nodes, node 3 and 34 respectively and node 34 has the maximum degree, our method finally divided node 10into community where the neighbors node 34 lies to, and the actual division is correct.

When the modularity Q is 0.4304, the corresponding community detection of our method is best. As can be seen from the figure 5, the Zachary network is divided two communities by the algorithm OCNS. White in the figure represents the center community with node 1 while grey represent the center community with node 34. Node 6,7,17 where the small community lies in is finally divided into the big community grey represents. It also shows that between the three nodes is the most closely associated. Node 9, 29, 31 with black labeled are both in the white and grey community. That is the

overlapping nodes obtained using the paper's algorithm. The degree of the node 1 and 34 is relatively big and similar to many other nodes, which explains the position of their respective community is self-evident. Because of this influence, the community structure formed has obvious core.

In order to further show the effectiveness of the algorithm, the method proposed compared with the 6 algorithms mentioned above. Figure 6 gives comparison of these algorithms modularity value, Table 2 gives the number of community and overlapping nodes.

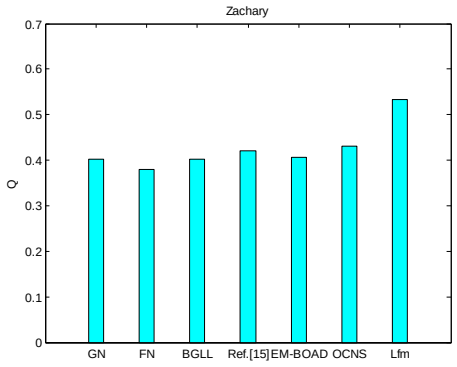

Fig. 6. Modularity comparison for the several algorithms (Zachary)

Table 2. The performance of the algorithms(Zachary)

algorithms	Q	The number of C	Overlapping nodes
GN	0.4013	5	0
FN	0.3806	3	0
BGLL	0.4020	3	0
Ref.[15]	0.4214	4	10
EM-BOAD	0.4060	2	3
OCNS	0.4304	2	9,29,31
Lfm	0.5333	2	3,9,10,14,31

As shown in Figure 6, the modularity value proposed in the paper is better than the first 5 algorithm, larger than algorithm GN, FN, reference [15] and [16], specific modularity value are given in table 2. Modularity is the quality standards to measure the community structure. The greater the modularity value, the better the results of classification. The results show that the algorithm is effective for the division of community structure. The algorithm gets two communities, in line with the actual situation of Zachary network. However, our algorithm modularity value less than Lfm algorithm, But compared to the above algorithm, overlapping nodes obtained more closely to the fact and node 3 and node 10 can return to the correct association.

4.2. Dolphin network

This data set is taken from the social network of 62 dolphin living in Doubtful Sound, New Zealand, and was compiled by Lusseau [20]. It describes the associations between dolphin pairs being the statistically significant frequent association.

Relevant similarity nodes of each node in Dolphin is as shown in figure 7:

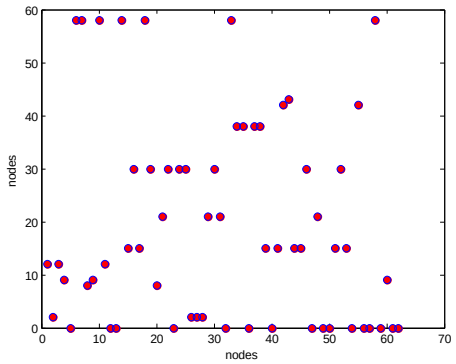

Fig. 7. Relevant similarity nodes of each node (Dolphin)

As can be seen from the figure 7. the similarity value between many nodes and other nodes is 0. If these nodes are directly involved in the community Division, it will affect the quality of community detecting, such as node 36 can be divided into other community. Therefore our algorithm firstly should find out these nodes, and finally to determine their classification.

The result of community detection in Dolphin is shown in figure 8.

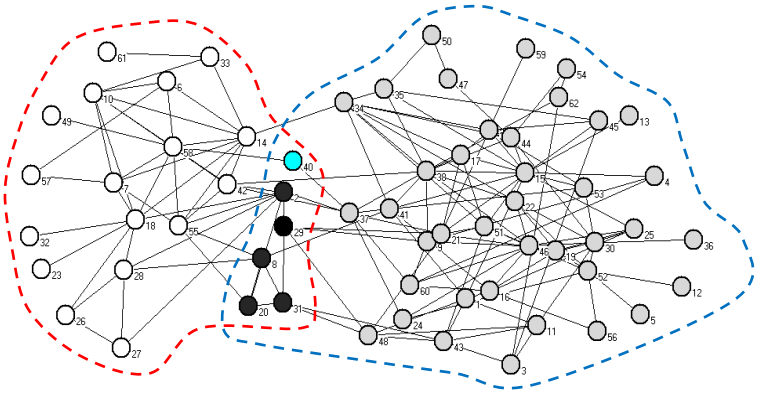

Fig. 8. Community structure of Dolphin network, Two communities are detected, labeled by white and grey circle, and overlapping part labeled by black circle, bridge node labeled by blue circle.

When the modularity value Q = 0.5480, community classification result is the best. Algorithm OCNS divides dolphin social network into two networks, labeled by white

and grey circle respectively. Nodes labeled by black circle is overlapping part. Then node 40 is both overlapping node and bridge node. The network is naturally divided into two associations. The accuracy rate of division is 100%.

In order to further show the effectiveness of the algorithm, the method proposed also compared with the 6 algorithms mentioned above. As shown in figure 9 and table 3.

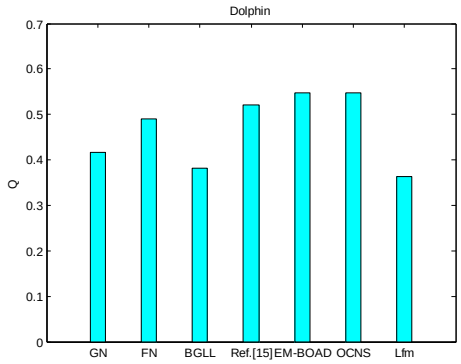

Fig. 9. Modularity comparison for the several algorithms (Dolphin)

Table 3. The performance of the algorithms(Dolphin)

algorithms	Q	The number of C	Overlapping nodes
GN	0.4155	5	0
FN	0.4897	3	0
BGLL	0.3816	2	0
Ref.[15]	0.5478	4	3,40,62
EM-BOAD	0.5210	3	9,40
OCNS	0.5480	2	2,8,20,29,31,40
Lfm	0.3632	2	8,20,29,31,40

As can be seen from figure 8, the modularity *Q* is the best in the paper, which is to say the division result is good and accurate. The specific modularity value and overlapping part are given in table 3. The overlapping nodes we get are consistent to the algorithm *Lfm*. It is more realistic than the Ref. [15] and [16]. All algorithms detect the node 40 as overlapping node, which indicate node 40 important in the actual network. In this paper, the node 40 is not only overlapping node, but also a bridge node, which is an important hub node between the two communities.

5. Conclusion

This paper presents an algorithm for mining the overlapping community based on node similarity. Algorithm OCNS's idea is simple and detects the overlapping nodes in the network after finite iteration. The experiment results proves the algorithm is not only effective and feasible but also more accurately reflect the real situation of the network.

Although the results are better than some existing classic algorithms, but there are still issues which need further study and discussion. For example, Jaccard correlation coefficients can only get the local information, not fully consider the topological structure of the whole network and we only consider the similarity between the nodes, without consideration of other information in the network topology.

Next, we will further study the weight of node and importance of node impact on community detection in the complex network. Whether or no, Community detection still has many problems worthy of studying, especially designing more effective overlapping community detection algorithm.

Acknowledgement. This work was sponsored by the Natural Science Foundation of Hunan Province, China (14JJ3062), Postdoctoral Science Foundation, China (2015M571150) and Natural Science Foundation of Hubei Province, China (No.2013CFB003).

References

1. Girvan M., Newman, M. E. J.: Community structure in social and biological networks Proceedings of the National Academy of Sciences. 99(12). pp7821-7826. (2002)
2. Kernighan, B. W., Lin S.: A efficient heuristic procedure for partitioning graghs. Bell System Technical Journal, 49(2): 291-301. (1970)
3. Fiedler, M.: Algebraic connectivity of graphs. Czechos lovak Mathematical Journal, 23(2): 298-305. (1973)
4. Girvan, M., Newman M. E. J.: Community structure in social and biological network, Proc. Natl. Acad. Sci, USA99 7821-7826. (2002)
5. Wu, Z. H., Lin, Y. F., Wan, H. Y., Tian, S. F., Hu K. Y.: Efficient overlapping community detection in huge real-world networks, Statistical Mechanics and its Applications, Vol. 391, 2475-2490. (2012)
6. Nepusz, T., Petroczi, A., Negyessy, L., Bazso, F.: Fuzzy communities and the concept of bridgesness in complex networks, Phys. Rev. E77(1) 016107. (2008)
7. Li, Y. P., Ye, Y. M., Wang, E. K.: Fast computation of modularity in agglomerative clustering methods for community discovery, Int. J. Adv. Comput. technol, 3(4),153-164. (2011)
8. Clauset, A., Newman, M. E. J, Moore, C.: Finding community structure in very large networks. Physical Review E, 70: 06611. (2004)
9. Blondel, V. D., Guillaume, J. L., Lambiotte, R.: Fast unfolding of community hierarchies in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10: 10008. (2008)
10. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435: 814-818. (2005)
11. Gregory, S.: A fast algorithm to find overlapping communities in networks. In Proceedings of the 12th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Antwerp, Belgium. (2008)

12. Lancichinetti, A., Fortunato, S., Kertrsz, J.: Detecting the overlapping and hierarchical community structure of complex networks. New J. Phys, 11: 033015. (2009)
13. Chen, D. B., et al.: An efficient algorithm for overlapping community detection in complex networks. Global Congress on Intelligent Systems, 68: 244-247. (2009)
14. Jaccard, P.: Etude comparative de la distribution floraledansune portion des Alpeset des Jura, Bullletin de la Societe Vaudoise des Sciences Naturelles 37, 547-579. (1901)
15. Chen, D. B., Shang, M. S., Lv, Z. H., Fu, Y.: Detecting overlapping communities of weighted networks via a local algorithm, Statistical Mechanics and its Applications, Vol. 389, 4177-4187. (2010)
16. Li, J. Q., Wang, X. Y., et al.: Detecting overlapping communities by seed community in weighted complex networks, Statistical Mechanics and its Applications, Vol. 392, 6125-6134. (2013)
17. Zachary, W. W.: An information how model for conflict and fission in small groups, J. Anthropol. Res, 33 452-473. (1977)
18. Qi, X. Q., Tang, W. L.: Optimal local community detection in social network based on density drop of subgraphs. Pattern Recognition Letters, 36, 46-53. (2014)
19. Lu, H., Wei, H.: Detection of community structure in networks based on community coefficients. Statistical Mechanics and its Applications, Vol. 391, 6156-6164. (2012)
20. Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., Dawson, S. M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol, 396-405. (2003)

Zuo Chen received his PhD degree in Hunan university in 2008. He is an assistant professor at College of Computer Science and Electronic Engineering, Hunan University. He is currently a postdoctoral of IIE CAS. His research interests include wireless sensor network, data mining and so on.

Mengyuan Jia was born on October 1989, in Henan Province, China. He is a postgraduate student of College of Computer Science and Electronic Engineering, Hunan University. His research interest is network security and data mining.

Bing Yang is a PhD and currently an assistant professor at Education Institute, Hubei University. His research interests is wireless sensor network.

Xiaodong Li is an assistant professor at College of Environment Science & Engineering, Hunan University. His research interests is Environment System Engineering.

Received: October 21, 2014; Accepted: April 28, 2015.