http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (11): 2160-2166, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.2160.2166

Parallel Overlapping Community Fast Detection Model Based on Local Modularity

Min Xu, Lijuan Zhou and Hui Liu
Information Engineering College, Capital Normal University, Beijing, 100037, China

Abstract: There are many communities in the complex network usually and the communities having these overlapping nodes in many occasion. How to detect community structures in complex network with quick and accurate methods become a hot topic in computer science, system science and social science and other related fields. This study makes an intensive study in the complex network and proposes a practical parallel processing algorithm which is based on local modularity to find the overlapping community structures in complex network. In order to test the accuracy and efficiency of the algorithm, a series of experiments have done in different real-world networks and random networks. These experiment data results show that the proposed algorithm can detect the community structures which are overlapping in limited time and work well in practical cases.

Key words: Overlapping community, community detection, complex network, parallel computing

INTRODUCTION

The concept of "Gemeinschaft" was brought up at (Community and Civil Society) published by German Sociologist (Tennis, 2001), expressing a form of organization based on the relationship of cooperation and it was translated as "community" later. During the past 100 years, the sociologists tried to define and research the community structures from different angles. As a result, the intension and extension of concept of community had changed a great deal compared with Tennis time.

With the development and popularization of internet, the concept of virtual community was extend based on community, its meaning is society bunching phenomena based on internet, that is, the formed individual relationship network on a certain degree scale of group. The space of virtual community is different from real community; the users in virtual community from different areas can express their opinion, direct interact and exchange their idea, as a result, the user can find his friends with common interests, then a group with cohesiveness can be formed.

In most cases, the community is not obvious but is hidden behind a great deal of complicated relationship. Researchers can find the community structures using the related method and technology and provide the personalized application and service based on the individual community relationship.

In fact many systems can be described by network which is consisted of the nodes set and the links set joining the nodes (Newman, 2004a). The vast amounts of data show the complex network is heterogeneous generally, that is to say, the complex network is composed by nodes which has different attribute. There is more relation between the same-type nodes, so the links between the same-type nodes are more than the different-type nodes. The sub graph which is composed of the same-type nodes and the links between these nodes is called as the community (Girvan and Newman, 2002). Interestingly, we find a wide variety of research about community detection in different areas and generally, the complexity of these algorithms of community detection is high. The paper describes one fast community detection algorithm which is parallel, concision and frank.

RELATED WORK

The research about community detection in complex network has been drawing more and more attention from researchers, many researched result about that subject has been published in domestic and foreign technical and scientific journals. The reprehensive research results are graph segmentation based on graph theory, hierarchical clustering algorithm, GN algorithm and W-H algorithm based on sociology and so on.

Graph segmentation algorithms: In general graph segmentation method can be divided into two classes as and K-L algorithm (Kernighan and Lin, 1970) which uses the bisection method to separate an entire graph to two optimal sub graphs at first, then separating each sub graph into two optimal parts iterative until the numbers of sub graph met the demand. In real application, the numbers of sub graph cannot be defined in advance and the scale of every community isn't quite same. This kind

Corresponding Author: Min Xu, Information Engineering College, Capital Normal University, Beijing, 100037, China

algorithm just fit for community detection in some community with special structure but normal complicate network.

Hierarchical clustering algorithms: How many communities are there in one complex network? How many nodes are there in every community? We haven't any idea before community detection. For these reasons graph segmentation as a general method to detect community is an unrealistic. Sociologists delivered agglomerative and divisive two methods about community detection based on the layer clustering method (Clauset *et al*., 2008). Agglomerative will compute the similarity degree of the two nodes according to the topology structure of network at first, then adding the edge to the graph which contains the nodes only but any edges as the similarity order from low to high repeatedly. The procedure can stop at any time; the number of community will be reduced gradually along with the clustering procedure.

Community dendrogram graph through clustering have not any correlation with the previous network topology, it just express the community structure of the previous network. Divisive method compute the similarity between each pair nodes based on the network topology firstly, then deleting the edges that connect two nodes with low similarity gradually until meeting the requirement. All the connected graphs in network are community at that time. Hierarchical clustering is hard to satisfy in practical applications, there are two reasons, firstly realistic network is generally sparse and a large amount of nodes don't belong to any community, secondly it is uncertain that how many communities should the network be divided .

GN algorithm and improved algorithms: Newman and Girvan (2004) both support the idea that the goal of the community detection is to group the nodes, after grouping, the nodes in one group have more links and the nodes in different group have less links. They proposed the GN algorithm of community detection what named by first letter of their name, the concept of betweeness was first introduced which is a extend concept of betweeness, it stands the number of the edge which belongs to all of the shortest path in the network (Freeman, 1977).

GN algorithm computes each edge-betweenness at first and then deletes the edge which has the biggest betweenness. The procedure divides the network to more and more community continuously until we get the required numbers of community. The algorithm supports the idea that the little edges in network are bottlenecks of communication between communities and they are in the path of community communication definitely. So, we consider how to find these edges which are the passage of the different communities. We will get the most nature decompound of the network if we delete these edges which have the top n highest betweenness.

The accuracy of GN algorithm is very high, at the same time, the algorithm is time consuming and low efficiency; In addition, on node only belongs to one community, it is not confident with the fact. There are many knowledge nodes belongs to the inter-discipline in group aggregation knowledge map, apparently, they should belong to several knowledge communities. Many researchers proposed new improved algorithm on the basis of GN algorithm, such as, Brandes *et al*. (2003) proposed a rapidly compute method, Golder *et al*. (2007) introduce the guide rule of how to find community structure automatically; Zhuge *et al*. (2006) proposed the weighted GN algorithm, Radicchi *et al*. (2004) give the definition of the strong community and weak community through the quantification method because they think it lacks the community definition of quantification in GN algorithm; They proposed the include self GN algorithm and Radicchi algorithm on the basis of above definitions (Radicchi *et al*., 2004).

K-clique community model and its community detection algorithms: The final goal of all these methods is to divide the network to several independent communities. While the community structure of the realistic network isn't independent completely, there are many overlap part in the two or more communities. For example, community detection as a knowledge node belongs to sociological area, at the same time it belongs to the field of computer science. For these reasons, Palla proposed the definition of K-clique community model and the algorithm to detect the community based on K-clique method. K-clique community model allow that each member can be indirectly connected but they can arrive any member in K steps (Palla *et al*., 2005). Its main idea is to get the shortest path matrix through the relationship between the nodes in the network structure at first, then create the K-neighbors matrix using the shortest-path matrix, finally detect the community structure in the network by the k-neighbors matrix.

Local community detection algorithms: It is unrealistic detecting the community by the global community detection algorithm in gigantic and dynamic evolving networks such as World Wide Web. Some local community detection algorithms were proposed, the most representative methods are the hub algorithm based on the idea of network hub nodes which is proposed by COSTA and BB algorithm proposed by BAGROW and BOLLT (Bagrow and Bollt, 2005). Hub algorithm support

the node which degree is the biggest is center node, what is called the hub node. The hub nodes have strong influence, so they will influent the neighbor nodes continuously. The first step of the algorithm is to select some nodes as hub nodes and then extend to other nodes which are met the requirement by centering the hub nodes. The limitation of the algorithm is that the number of the community has to be defined in advance. Another weakness of the algorithm is we need to know the entire network topology although we just use the local information of nodes during the algorithm processing. BB algorithm extends the D-BALL based on the hub algorithm; the algorithm finds the community structure of the node starting at any node through extending the D-BALL. Whether the algorithm is correct is closely related to the position of the node in the entire community. If the given node is just the center node of the network, the detected community will be right. While the given node is the edge position node, the algorithm will detect the community which includes the realistic two communities. So we think the result of algorithm is hard to express the real condition.

Parallel overlapped community fast finding model: The tradition method of how to find community structures in complex network is to select the nodes which have the biggest characteristic value at first, then to search among the neighbor nodes centered by these nodes through the various directions. The procedure takes the nodes which met the requirement into the community continuously until the community reaches the extreme value condition. All above action will repeat in the remaining nodes which are not be included in any community and the algorithm will end when all of the nodes will be include in one community.

Now these algorithms have two limitations in general, the first, one node only belongs to one community what is not fit for the many real situation; The second, computational complexity of these algorithm is o(n3) generally and n stands the number of nodes in complex network:

$$Q = \sum_{i=1}^{p} (x_{ii} - \sum_{j} x_{ij}2) \quad (1)$$

So, the paper proposes a new fast parallel overlapping community detection algorithm inspired by the local modularity because of the limitation of these algorithms (Hereafter referred to as POCFA). Modularity is a guideline to measure the quality of community division developed by Newman (2004b). n. G (V, E) expresses one undirected complex network graph including n nodes. If we divide the network into the number of p communities by some ways, we can construct a symmetry matrix X with p rows and p columns. The equation expresses the computing method of Modularity Q, the matrix element xij denote the value of the number of all link in the network divided by the number of the link between community i and community j and modularity Q denote the number of all link in the network divided by the number of link among the local community, the value of modularity Q is from 0 to 1; its value is closer to 0, the community structure is less obvious; its value is closer to 1, the situation is opposite. Although the result of community structure based on the modularity is better, the algorithm need do a large amount of computation and time consuming, so clause proposed the concept of local modularity based on it.

In order to clearly describe the parallel overlapping community fast detection algorithm, this paper defines several related concepts as below:

- **Node contribution degree m:** Suppose node v belongs to the community, L_{in} denotes the number of link connecting node v and the other nodes in the community, L_{out} denotes the number of link connecting node v and the outer nodes which are not belong to the community. The following equation is how to compute the node contribution. If the value of m is higher, the node takes more contribution to the community:

$$m = \frac{L_{in}}{L_{in} + L_{out}} \quad (2)$$

- **Community Modularity M:** N in follow equation denote the number of nodes in community, $\bar{D}$ denote the average distance from the center node to other nodes in local community, m equals the number of link within the community adding the number of link connecting the community and the outer part in network, then divided by the number of link within the community:

$$M = N * 1/\bar{D} * \frac{L_{in}}{L_{in} + L_{out}} \quad (3)$$

- **Community overlapping degree O:** In fact, the community overlapping degree of the two communities is the proportion of the number of common nodes in two communities and the number of all nodes of the two communities. Supposed the overlapping degree is greater than 0.6 or one threshold, we think the two communities is high overlapping. In that case, we can take into account merging the two communities into one community:

$$O = \frac{C_A \cap C_B}{C_A \cup C_B} \quad (4)$$

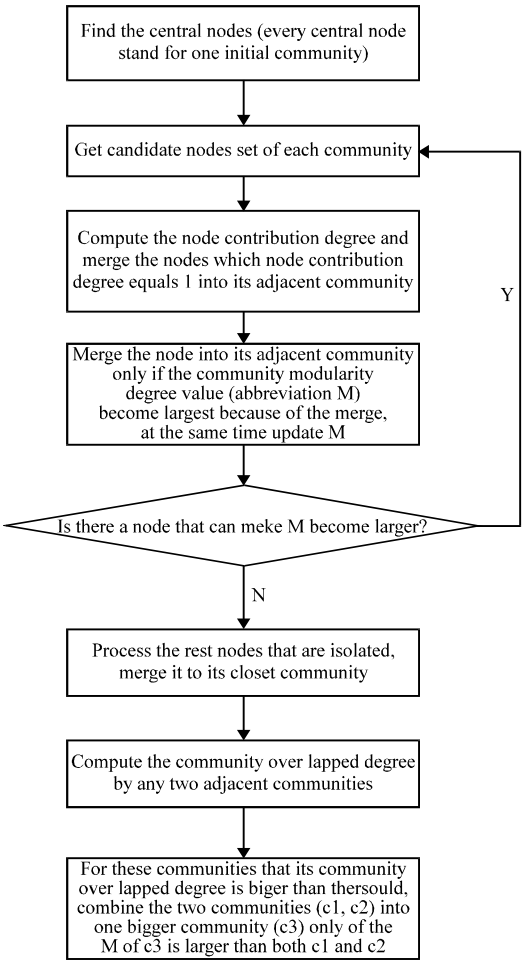

Fig. 1: POCFA algorithm processing step

The algorithm description of the Parallel Overlapping Community Finding Algorithm Based on Local Modularity is as below (algorithm processing step is shown in Fig.1):

Step 1: The center nodes with the biggest degree are taken as hub nodes of the initial communities, at that time, there are just one node that is the hub node in each community and the initial value of community modularity M is 0

Step 2: Take the adjacent nodes of the local community into the candidate nodes set

Step 3: Compute the node contribution degree value of every node in the candidate nodes set, if the node contribution degree value equals 1, the merge flag of the node is set as 1

Step 4: All nodes which merge flag is 1 will be merged into the community. After finishing the merge operation, the algorithm will compute the value of community modularity degree M. If none of the nodes with the merge flag 1, we will assume the node Vi belongs to the community again and compute the community modularity degree value. (It is called assumed community modularity degree.)If the maximum assumed community degree max(Mi) is greater than M, the node will be take into the community and update the M to max(Mi);otherwise, turn to step 6

Step 5: Repeat step 2, 3 and 4

Step 6: All the community structure has been detected

Step 7: All above the six steps is executed at the same time, that is to say, the number of community detection algorithm function is that the number of the hub nodes

Step 8: If there are nodes which are not be merged into any community after the parallel procedures were done, then compute their node contribution degree in each community and merge the node to the community in which the node has the greatest node contribution value

Step 9: When all the communities are detected, the algorithm will compute the community overlapping degree between any two communities, if the overlapping degree is greater than the threshold, the two communities will be merged into one community until the overlapping degree of the any two communities is less than threshold

EXPERIMENTS

Community detection in a simple cognitive network: We detect the community structure in the graph using the POCFA. There are 16 nodes in the graph (show as Fig. 2), we get and the 3 hub nodes. POCFA parallel expand the 3 communities which is centered by one hub node. Community A with hub node n4, community B with hub node n11 and community C with hub node n16 expand side by side until their modularity degree reach the extreme value 25/6, 392/81, 16/5. At that time, community detection is done. The following work is to compute the community overlapping degree of any two communities, the overlapping community degree is all far less than the threshold, so the algorithm ends. Through the POCFA, we get the 3 cognitive communities (show as Fig. 3), as they are denoted as below graph. There is common node n6 in community A and community B; there are common nodes n12 and n13 in community B and community C. We also did the experiment and got the experiment result data with G-N algorithm and Factions algorithm. The experiment result is shown in Fig. 4 and 5. The detail comparison of the three experiment result is listed as Table 1.

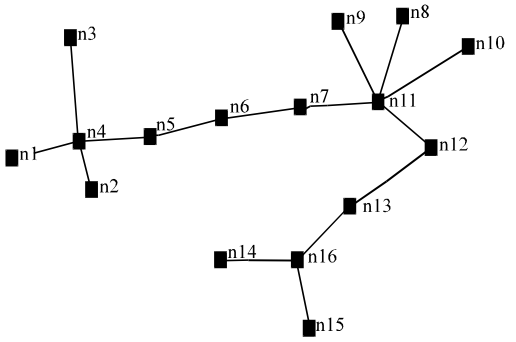

Fig. 2: Simple knowledge network

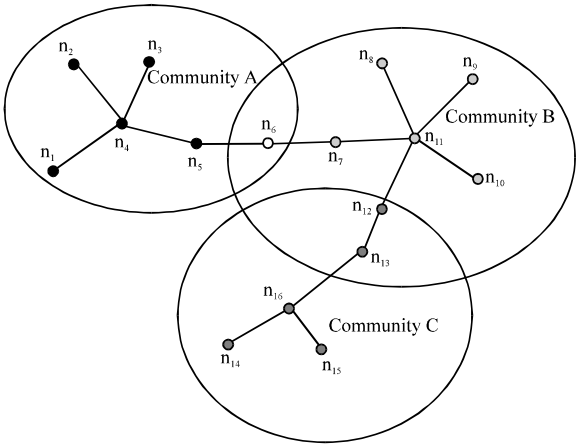

Fig. 3: Communities in simple knowledge network by POCFA algorithm

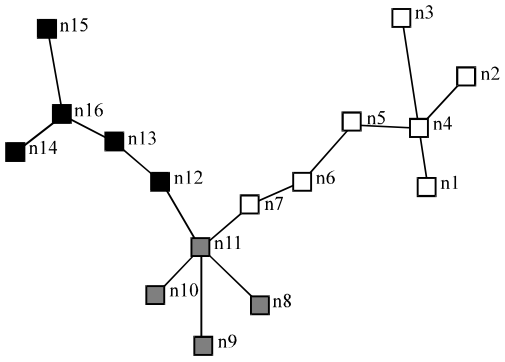

Fig. 4: Communities in simple knowledge network by G-N algorithm

Experiment about the classical dataset karate club social network: Karate club social network is a dataset describe the social relationship of the karate club according to the observation of Wayne Zachary during two years. Now it becomes a classical dataset in community detection area,

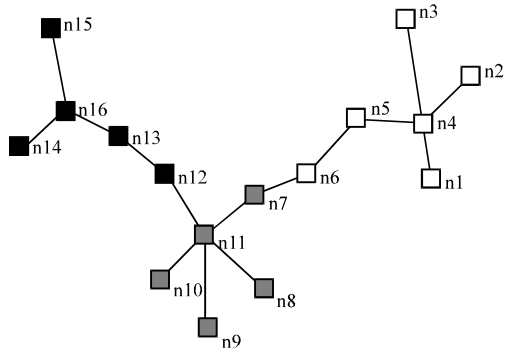

Fig. 5: Communities in simple knowledge network by factions algorithm

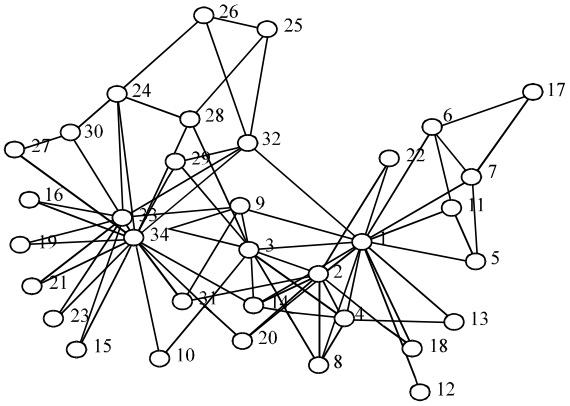

Fig. 6: Karate club social network (before division)

Table 1: Comparison of the three algorithm of community detection in simple knowledge network

Method Algorithm	Community A	Community B	Community C
POCFA	n1, n2, n3, n4, n5, n6	n7, n8, n9, n10, n11, n12	n13, n14, n15, n16
G-N	n1, n2, n3, n4, n5, n6, n7	n8, n9, n10, n11	n12, n13, n14, n15, n16
Factions	n1, n2, n3, n4, n5, n6	n7, n8, n9, n10, n11	n12, n13, n14, n15, n16

many related new research will verify their algorithm and other work using that dataset. There are 34 members in Zachary network (shown as Fig. 6), but the club was divided as two little club (shown as Fig. 7) because of the terrible breach between the university president and club executive.

By POCFA, we first select n34 and n1 as hub nodes, algorithm take them as central node of each community, the community which central node is n34 successively aggregate n10, n15, n16, n33, n19, n21, n23, n30, n27, n24, n28, n31 and n9. After n9 is aggregated into the community, the M value of this community reaches the maximum 0.666667; the community which central node is

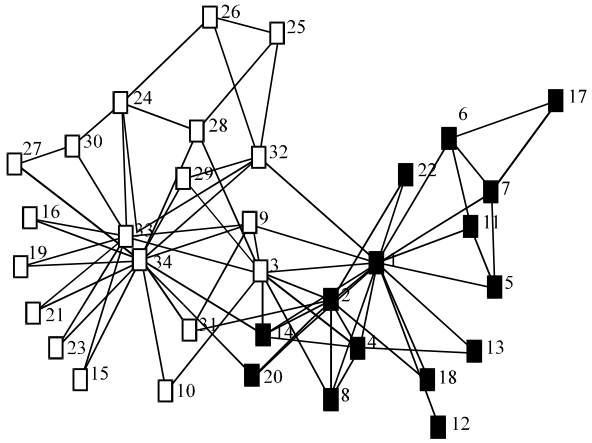

Fig. 7: Karate club social network (after division into two clubs)

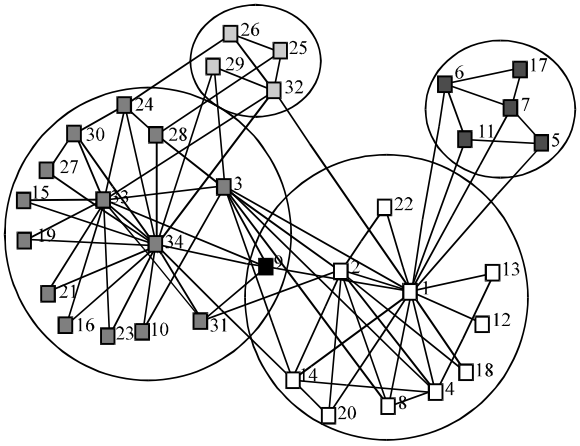

Fig. 8: Initial community detection result in karate club social network by POCFA

n1 successively aggregate n12, n13, n4, n8, n2, n14, n3, n18, n22, n20 and n99, after n99 is aggregated into the community, the M value of this community reaches the maximum 0.625. After the algorithm iterates one time, there are still some nodes which are not aggregated into any community. The n32 and n7 are taken as central node of two new communities, the community which central node is n32 successively aggregate n25, n26 and n29. After n29 is aggregated into the community, the M value of this community reaches the maximum 0. 363636; the community which central node is n7 successively aggregate n17, n6, n5 and n11, after n11 is aggregated into the community, the M value of this community reaches the maximum 0.48. Through the algorithm two iterations, these all nodes have been belonged to at least one community, that is to say, community initial detection has finished. Its result is that there are four communities that are shown in Fig. 8.

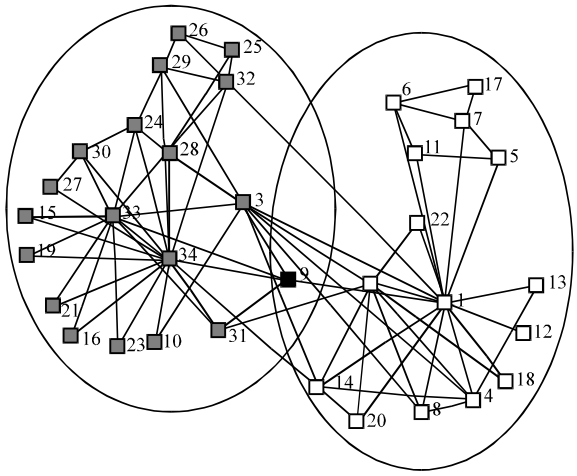

Fig. 9: Community detection result in karate club social network after combination by POCFA

Table 2: Comparison of the three algorithm of community detection in karate club social network

Method Algorithm	Community 1	Community 2
Real situation	1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22	3, 9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
POCFA	1, 2, 3, 4, 5, 6, 7, 14, 17, 18, 20, 22	9, 10, 15, 16, 19, 21, 23, 24, 8, 9, 11, 12, 13, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
Factions	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 17, 18, 20, 22	9, 15, 16, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
G-N	1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22	3, 9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34

The next step is to try to combine the communities that the overlapping degree is larger than threshold value 0.7. In the previous work, we have found four communities £¬that are community 1 (hub = n34), community 2 (hub = n1), community 3 (hub = n32), community 4 (hub = n7). After combination, the M value of the new community which is aggregated by community 1(the M values is 0.666667) and community 3(the M values is 0.363636) is 0.736842; the M value of the new community which is aggregated by community 2(the M values is 0.625) and community 4(the M values is 0.363636) is 0. 76087. The M values of the new communities are larger than the previous part community, so we process the combination .The detail communities graph is shown in Fig. 9.

The detail algorithm comparison experiment result between POCFA algorithm and other two algorithms about Zachary classic dataset is listed in Table 2, it shows that the POCFA algorithm is very efficient and the correct rate of community detection exceed 95%, in addition, the POCFA algorithm also find the node 9 is the common node of the two club communities.

CONCLUSION

This study discusses the definition of community detection and then reviews some algorithms, method and technology to detect the community structure. As we all know, Identifying overlapping communities in networks is a really challenging task. This study propose a new parallel overlapping community fast detection algorithm based on local modularity, the algorithm concurrent expands the every community centered by the hub node according to the extreme value of the local modularity until finding the entire community and then processing the community aggregation according to the community overlapping degree. The algorithm has the advantage of parallel processing in the first step of community initial detection, so it can improve the algorithm efficiency. We make a group of experiments and a thorough comparison of POCFA and other two classical algorithms is provided.

ACKNOWLEDGMENT

This study is supported by Scientific Research Common Program of Beijing Municipal Commission of Education (Grant No. KM201110028018).

This Study is supported by the National Natural Science Foundation of China (Grant No. 31101078) and Beijing Natural Science Foundation of China (Grant No. 4112013).

REFERENCES

Bagrow, J.P. and E.M. Bollt, 2005. Local method for detecting communities. Phys. Rev. E, Vol. 72, 10.1103/PhysRevE.72.046108

Brandes, U., M. Gaertler and D. Wagner, 2003. Experiments on graph clustering algorithms. Proceedings of 11th European Symposium on Algorithms, September 16-19, 2003, Budapest, pp: 568-579.

Clauset, A., C. Moore and M.E. Newman, 2008. Hierarchical structure and the prediction of missing links in networks. Nature, 453: 98-101.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. Sociometry, 40: 35-41.

Girvan, M. and M.E.J. Newman, 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA., 99: 7821-7826.

Golder, S.A., D.M. Wilkinson and B.A. Huberman, 2007. Rhythms of social interaction: Messaging within a massive online network. Proceedings of the 3rd International Conference on Communities and Technologies, June 28-30, 2007, State University, East Lansing, Michigan, pp: 41-66.

Kernighan, B.W. and S. Lin, 1970. An e?icient heuristic procedure for partitioning graphs. Bell Syst. Tech. J., 49: 291-307.

Newman, M.E. and M. Girvan, 2004. Finding and evaluating community structure in networks. Phys. Revi. E, Vol. 69, 10.1103/PhysRevE.69.026113

Newman, M.E.J., 2004a. Detecting community structure in networks. Eur. Phys. J. B: Condens. Matter Complex Syst., 38: 321-330.

Newman, M.E.J., 2004b. Fast algorithm for detecting community structure in networks. Phys. Rev. E, Vol. 69. 10.1103/PhysRevE.69.066133

Palla, G., Derenyi, I., I. Farkas and T. Vicsek, 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435: 814-818.

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto and D. Parisi, 2004. Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA., 101: 2658-2663.

Tennis, F., 2001. Community and Civil Society. Cambridge University Press, UK.

Zhuge, H., L. Ding and X. Li, 2006. Networking scientific resources in the knowledge grid environment. Concurrency Comput. Pract. Experience, 19: 1087-1113.