ATLANTIS PRESS

Advances in Engineering Research, volume 155

2nd International Conference on Advances in Energy, Environment and Chemical Science (AEECS 2018)

Entropy-based Social Network Link Partition Algorithm

Shusen Zhang[1, a], Xun Liang[1, b, *], Xiaoping Zhou[1, 2, c], Xuan Zhang[1, d]

[1]School of Information, Renmin University of China, Beijing, China

[2]School of Electrical & Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China

[a]zss2446@ruc.edu.cnl, [b]xLiang@ruc.edu.cn, [c]pingz@ruc.edu.cn, [d]xuanz@ruc.edu.cn

*Corresponding author

Keywords: Social computing, Social networks, Community division, ELP algorithm

Abstract. It is an important core issue in social networks to divide community or group. And, the network node is the mainstream of community division algorithm as the processing object to divide the network. This paper introduces the entropy theory into social networks partition on the basic of studying the concept of entropy and social networks partition algorithm. And we proposed an Entropy-based Link Partition algorithm (ELP algorithm), which is the social network links as the processing object. Also, the similarity between two objects is properly defined and improved, which thus is more close to the real situation of the social network. Experimentation on two real-world networks, and we obtained results of community division and compared with other community partition algorithms to verify the effectiveness of the proposed algorithm. The ELP algorithm has a higher accuracy, and communities are more realistic than that generated by either of the Link Clustering algorithm (LC) or the classical Clique Percolation Method (CMP).

1. Introduction

Social networking is a relational structure formed by the connection between users and users, and it is an extension of the user's real social relationships on the network. The rise of the social networks concept from its proper description of social interaction, and proposed by Simmel Georg, a German sociologist, from the view of sociology [1]. Indeed, it is very early that people began study the social networks, and people have been hoping to find inner mechanism of social relations, evolution and the propagation pattern of information in the social network, etc. Social networks exist in all areas of life, which affect people's work, study and all aspects of life, and it is very necessary to study social networks. Especially the online social networks, hundreds of millions of people on the Internet every day to communicate, work, study and other activities, and people working and living in a real "virtual" space, thus forming a virtual social networks. And the emergence of Face book, Twitter, Micro-blog, blog and other online social network applications extends the real world of human beings in the virtual world of the network. It can be said that it establish a new type of social relations between people, and gradually changing the traditional face to face communication. In order to analyze this social network, the current commonly used method is to extract important features in networks. This means converting a larger network to a smaller network, which is an effective summary of the larger network, and remains the important features of the original network [2]. In social networks, this approach can be achieved through two ways, one is to identify users' groups, that is the community division or discovery. The other is to identify nodes that have the same structure, position, effect, or play the same role in networks, namely the role identification.

Social networks can be regarded as consisting of many communities which with strong homogeneity or connection between nodes within the network community, while heterogeneous, or sparsely connected among different communities. In fact, the social network partition (community division) has long been a hot spot in social network analysis. The complexity of the social network structure brings great difficulties to researchers, and through the research of community can greatly

reduce the study difficulty. Dividing social networks into reasonable communities, and it is very important for us to understand the structure, characteristics, functions and evolution of the whole community and to make full use of the network value. In general, it is of great significance for us to understand the whole network structure and analyze the characteristics of the network by dividing the community in the network.

In this paper, we first analyze the community partition algorithm and introduce the concept of entropy into social networks partition, and properly define and improve the similarity between the two objects. Then, we propose an Entropy-based Link Partition algorithm (ELP algorithm), which is the social network links as the processing subject. Finally, experimentation on real-world networks, we obtained the results of community division and compared with link clustering algorithm (LC algorithm) and CPM algorithm. We have come to the conclusion that the proposed algorithm in this paper has a higher accuracy and enables efficient divide communities.

2. Related Works

Because of the importance of community division, researchers have proposed a variety of community division methods. Initially, the researchers through the graph partition method to divide community, and mainly obtain approximate solutions of the community structure by some tentative algorithm. For example, in 1970, Kernighan and Lin [3] proposed the famous Kernighan-Lin algorithm. Kernighan-Lin algorithm is a heuristic optimization dichotomy algorithm based on the greedy strategy, which continually divides the network into two sub networks of known sizes. However, the algorithm can only get the local optimal solution， and it is sensitive to the initial solution. Fiedler M et al [4-5] proposed a spectral bisection algorithm based on the laplace eigenvectors of graphs. In this algorithm, the similarity between nodes is used as the weight of edges, and it gets an undirected weighted graph, and then graphically divide it. With similar to the Kernighan-Lin algorithm, the spectral bisection algorithm also divides the network into two sub networks. However, when the number of communities within the network is uncertain, the algorithm is not ideal. To deal with the problem of overlapping communities, Palla [6] et al proposed the clique percolation method (CPM), which is subsequently widely used for overlapping community structure detection. And Wu and Huberman [7] proposed the Wu-Huberman rapid spectral segmentation method, etc. Based on the hierarchical clustering method, Girvan and Newman [8] proposed the famous GN algorithm. The algorithm divides the graph into smaller units by constantly calculating the link betweenness and removes edges to get the reasonable results of community division. However, the algorithm time complexity is higher, and is generally suitable for dealing with a smaller scale of social networks. Flake [9] et al proposed the MFC algorithm based on the max flow/min cut theory in graph theory. Based on greedy strategy, Newman [10] proposed a Newman fast algorithm by maximizing the modularity evaluation function (Q function). On this basis, Newman [11] et al also used the heap to calculate and update the modularity of the obtained community, and proposed CNM algorithm. In addition, Aaron Clauset [12] proposed a community partition algorithm based on the local modularity. Fortunato [13] et al proposed a community discovery algorithm based on local optimal function. In order to comprehensively consider the influence of the node attribute and the topology structure of the network, and make the community structure more optimized. Hong Cheng [14] et al proposed a SA-cluster graph clustering algorithm based on structure and attribute. Xiaowei Xu [15] et al proposed the SCAN algorithm, and Youfang Lin [16] et al proposed CIG_ESC algorithm et al.

While community division methods in social networks have achieved great development, there are still facing many problems in the algorithm. For example, ignoring the effects of nodes attributes; the difficult to handle large-scale network data; the time complexity is high, et al. Therefore, the community division in social networks needs to do much more.

3. Entropy-based Social Networks Partition Algorithm

3.1 Entropy Clustering

According to the theory of information entropy proposed by Shannon: the more chaotic data, the greater the entropy value. We make the following assumption, the dataset $X = \{ x_1, x_2, x_3 \cdots x_n \}$ is n data points in M dimensional space, in which the entropy value of the *i*-th data point x_i is:

$$Ei = -\sum_{j \in x}^{j \neq i} (S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})) \tag{1}$$

Among them, S_{ij} is the similarity between two data points (*i*, *j*), which is calculated by Euclidean distance. Function $S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})$ can achieve maximum value 1 when $S_{ij} = 0.5$， and S_{ij} increasingly close to 0 or 1, the entropy value closes to 0, namely:

$$\begin{cases} \lim_{S_{ij} \to 0} (S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})) = 0 \\ \lim_{S_{ij} \to 1} (S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})) = 0 \end{cases} \tag{2}$$

Based on the above assumption, we can infer that data points that have little effect on the entropy of a data point (such as the data point x_i) are those that are very close or relatively far from the data point (x_i). And data points that have a greater effect on the entropy of a data point (such as the data point x_j) are those that approximate the average distance from this data point (x_j). After the analysis, we can draw the conclusion that those data points that have small entropy should be mainly located at the central position of the relative concentration of data points, and these data points are likely to be the cluster center. Those data points that have large entropy value should be mainly located at the boundary position of the relatively sparse of data points, and these data points are unlikely to be the cluster center. In general, the entropy-based clustering method is mainly through the calculation of entropy to determine the number of clusters and the cluster centers. The calculation of the entropy for each data point is based on the similarity measure. The data point with the smallest entropy value is selected as the cluster center. Among the remaining data points, those data points that whose similarity degree to the cluster center is greater than a threshold value β form a cluster. And then, these data points are removed from the data set, and repeat until the data set is empty.

3.2 Link Clustering

In 2010, Ahn et al [18] proposed the Link Clustering (LC) method, which is a relatively new method for detecting overlapping communities in networks. In community division methods, the traditional methods generally take the node as the research object, while LC method's research object is the link between nodes in networks. The basic propose of LC is to derive a transform matrix whose elements are composed of the link similarity. Clustering by hierarchical clustering method, and through the partition density on the resulting dendrogram to determine the cut level for best community division. [19] In link clustering, the link belongs to only one community, but for the node, there are many links to connect. Therefore, after transforming the links' community result to a nodes' community result, it will be appear that nodes belong to different communities. Thus, we naturally obtain the detection results of overlapping communities.

Link clustering algorithm measures the similarity based on the Jaccard distance. Assuming that the set of node *i* and its neighbors as $n_+(i)$, and e_{ij} represents the link between node *i* and node *j* in network. We define the adjacent link as two links with a common node, and the similarity LS1 between links e_{ik} and e_{kj} to be defined as:

$$LS1 = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \tag{3}$$

If there is no common nodes in two links, and the link similarity (LS1) is 0.

3.3 Entropy-based Link Partition Algorithm (ELP)

Based on the above, the link in the network as the research object in this paper. We introduce the concept of entropy into the social network partition, and propose an Entropy-based Link Partition algorithm (ELP algorithm). First, according to the actual situation of links, we calculate the similarity between links through an improved Jaccard distance. Then, we considering the links in the network as points in the Euclidean space, and calculate the entropy of these points based on the similarity between the links. Finally, according to the entropy clustering principle, we get the result of link clustering directly. In order to solve the problem that nodes belong to different communities, we determine based on the number of links, and these links contain the node and are in different communities. And, the community with the highest number of links is the one to which the node is finally divided. In addition, the clustering results based on entropy are data points that with obvious differences between clusters. Also, the dissimilarity between communities based on entropy clustering is also better. In fact, the definition of similarity has a great influence on the final result of the community division. In the formula (3), the calculation of the link similarity (LS1) is based on the Jaccard distance, and it can only calculate the similarity between adjacent links, while ignoring the similarity between non-adjacent links. And it will directly affect the analysis effectiveness of community structure.

In order to introduce and retain more information about links, we choose another improved similarity based on Jaccard distance, namely the similarity LS2 of two links e_{ij} and e_{mn} is:

$$LS2 = \frac{\left|n_+(i)\cap n_+(m)+n_+(i)\cap n_+(n)+n_+(j)\cap n_+(m)+n_+(j)\cap n_+(n)\right|}{\left|n_+(i)\cup n_+(m)+n_+(i)\cup n_+(n)+n_+(j)\cup n_+(m)+n_+(j)\cup n_+(n)\right|} \tag{4}$$

However, the similarity simply relies on the Jaccard distance, and it is not appropriate to the adjacent links. The adjacent links should have a higher similarity, and we redefined the similarity between two adjacent links as:

$$LS2 = \frac{\left|n_+(i)\cap n_+(m)+n_+(i)\cap n_+(n)+n_+(j)\cap n_+(m)+n_+(j)\cap n_+(n)\right|+2}{\left|n_+(i)\cup n_+(m)+n_+(i)\cup n_+(n)+n_+(j)\cup n_+(m)+n_+(j)\cup n_+(n)\right|+2} \tag{5}$$

Compared with the similarity LS1, the similarity LS2 can retain more link information for adding the information between non-adjacent links, and the similarity measure is also more close to the actual situation. In fact, this similarity is more consistent with the nature of friendships in real social networks. And those who have commons friends in the same community are more likely to be friends and share some common tendencies. In terms of time complexity, the ELP algorithm in this paper calculates the entropy of each edge, and obtains the final clustering result based on the obtained entropy and similarity. The time complexity is $O(n^2)$. Compared with the LC algorithm, although the time complexity is the same, the processing of the ELP algorithm is relatively simple and convenient.

4. Experiment

In order to verify the effectiveness of the ELP algorithm, we experiment in the standard data and compare the effect of community division under different algorithms and conditions.

4.1 Data Source

In experiment, we selected two accepted standard data sets (two real-world networks data sets) to experiment.

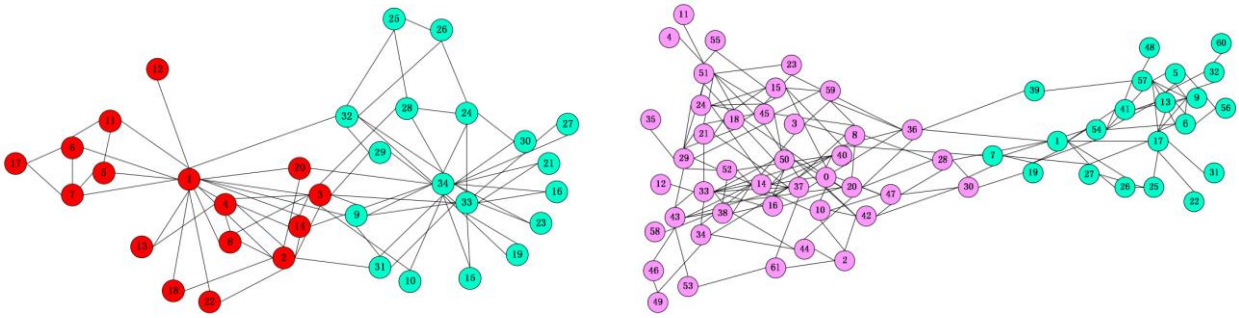
1) Zachary Karate Network

Karate network (Zachary's Karate Club) [20] is a social network of friendships between 34 members of a karate club at a US university. The scenario represented is that of a karate club being split into two new organizations as a result of a disagreement over pricing between club president

John A. (33) and instructor Mr. Hi (1), as shown in Fig. 1(a). The network has 2 reference classes with 34 nodes and 78 links.

2) Dolphins Social Network

Dolphin Social Network [21] is a relation network between bottlenose dolphins. Long-lasting associations are a strong feature of the community structure and this stability in the dynamics of association was observed within and between the sexes. The network has 2 reference classes with 62 nodes and 159 links, as shown in Fig. 1(b).



(a) Zachary Karate Network (b) Dolphins Social Network

Fig. 1. Two real-world networks.

4.2 Experimental results

1. LC Algorithm

We first divide the experimental network by LC algorithm. The karate network was divided into 22 communities, while the Dolphin social network was divided into 53 communities. We only showed the result of karate network in this paper, as shown in Fig. 2, and each color represents a community.

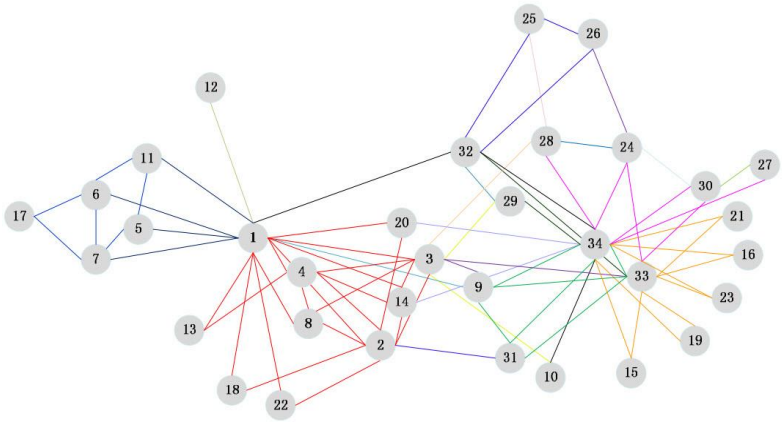


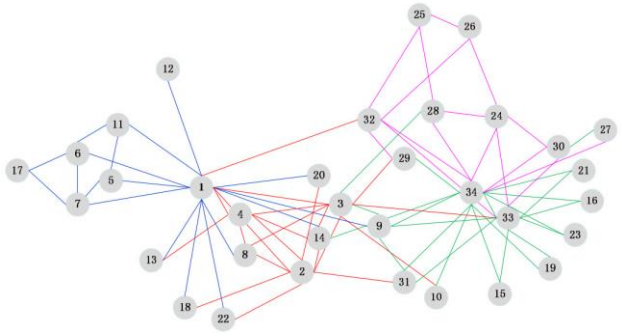
Fig. 2. LC algorithm partition result

From the above, the division effect of the LC algorithm in these two networks is not unsatisfactory and is quite different with the standard community division in Fig. 1.

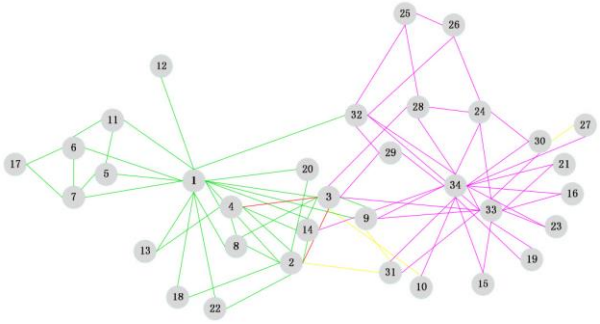
2. ELP Algorithm

We through ELP algorithm processed the experimental network, and observed the results of community division.

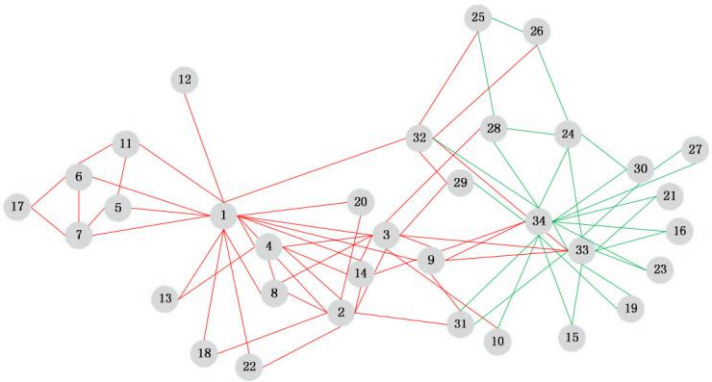
1) The community division result of the karate network under the ELP algorithm as shown in Fig. 3. In which, when similarity threshold is 0.1 (divided into 4 communities), the results as shown in Fig. 3(a), 0.05 (divided into 4communities) as shown in Fig. 3(b), and 0.01 (divided into 2 communities) as shown in Fig. 3(c).



(a) similarity = 0.1



(b) similarity = 0.05

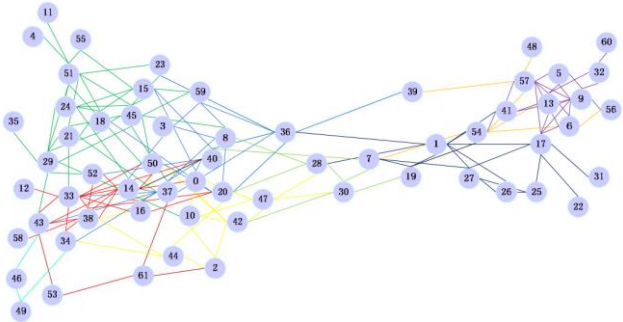


(c) similarity = 0.01

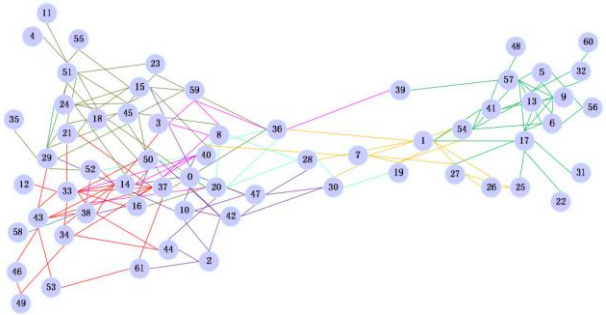
Fig. 3. ELP algorithm partition result

The community division effect of the ELP algorithm has a great relationship with the similarity threshold. As long as the similarity between objects is greater than the threshold value, these objects will be divided into a community. Therefore, with the decrease of the similarity threshold, the objects are more and more easy to be divided together, and the number of the community will be less and less. From the above experiment, the partition results of the ELP algorithm are significantly better than that of the LC algorithm. The ELP algorithm can effectively realize the community division, and is very consistent with the real situation when the similarity threshold is 0.05, namely, Fig. 3(b).

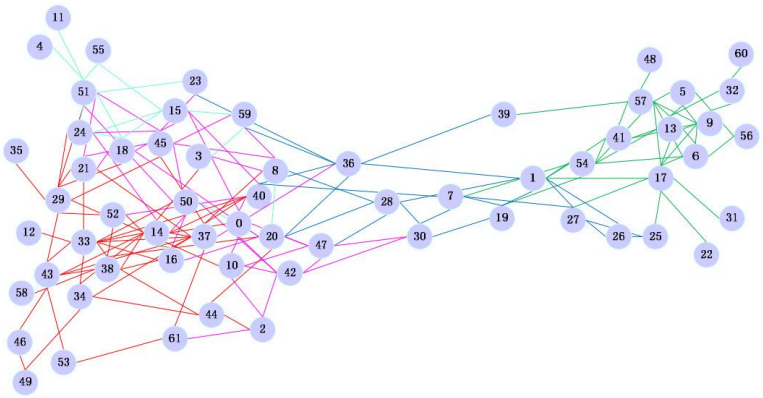
2）The community division results of the dolphin social network under the ELP algorithm as shown in Fig. 4. In which, when similar similarity threshold is 0.1 (divided into 11 communities), as shown in Fig. 4(a), 0.05 (divided into 8 communities) as shown in Fig. 4(b), and 0.01 (divided into 5 communities) as shown in Fig. 4(c).



（a）similarity = 0.1



(b) similarity = 0.05



(c) similarity = 0.01

Fig. 4. Partition result of dolphin social network in elp algorithm

From the above experiment, as the similarity threshold decrease, the community partition results of the ELP algorithm is getting better and better, more and more close to the true division of the real situation. And the ELP algorithm can effectively realize the community division in dolphin social network.

3. Accuracy Comparison

We also through CMP algorithm processed the experiment data sets, and got the results of community division. According to the experimental results, we compare the community division accuracy of the experiment data sets under LC algorithm, CPM algorithm and ELP algorithm, and the results as shown in Table 1, and the S represents similarity.

Table 1 The accuracy of the community division (%)

Algorithm / Data	LC	CPM	ELP		
			S=0.1	*S=0.05*	*S=0.01*
Zachary Karate Network	51	94	74	100	94.4
Dolphins Social Network	28.8	74	53	67.3	82

From the table 1, the ELP algorithm proposed this paper has a higher accuracy compared with LC algorithm and CPM algorithm.

By comparing the above results, we can see that the proposed algorithm can effectively realize community division, and it has a higher accuracy compared to other two division algorithms.

5. Discussion

In community division algorithms, most of them take the nodes of networks as the analysis object, and seldom take the links as the processing object. In this paper, we take the link in social networks as the research object, and proposed an Entropy-based Link Partition algorithm (ELP algorithm) based on studying entropy clustering process and link clustering algorithm. Also, the similarity between links is also properly defined and improved. The ELP algorithm can effectively realize the community division and has a higher accuracy compared with LC algorithm and CPM algorithm. Although this algorithm is helpful to analyze the network community structure, since the time complexity of the ELP algorithm is O (n2) and the limitation of similarity definition, this algorithm is suitable for handling small scale network data, but not for large-scale network data. In the further work, we can search for the definition of similarity between non-adjacent links in large scale network data, and study the solution of parallel processing for large scale network data through distributed platform.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71531012, 71601013); the Natural Science Foundation of Beijing (Grant No. 4172032).

References

[1] H. Rainie and B. Wellman, Networked: The New Social Operating System. London: The MIT Press, (2012).

[2] M. Forestier, A. Stavrianou, J. Velcin, and D. A. Zighed, "Roles in social networks: Methodologies and research issues," Web Intelligence and Agent Systems, vol. 10, no. 1, pp. 117–133, (2012).

[3] B.W. Kernighan and S. Lin, "An efficient heuristic Procedure for Partitioning graphs," Bell System Technical Journal, vol. 49, no. 2, pp. 291-307, (1970).

[4] M. Fiedler, "Algebraic connectivity of graphs, Czechoslovakian Mathematical Journal," vol. 23, no. 2, pp. 298-305, (1973).

[5] A. Pothen, H.D. Simon, and K.P. Liu, "Partitioning Sparse matrices with eigenvectors of graphs," SIAM J. Matrix Anal.APPl. vol. 11, no. 3, pp. 430-452, (1990).

[6] G. Palla, I.J. Farkas, P. Pollner, I. Derenyi and T. Vicsek, "Directed network modules," New Journal of Physics, vol. 9, no. 6, pp. 186-207, (2007).

[7] F Wu and B.A. Huberman, "Finding communities in linear time: a physics approach," The European Physical Journal B-Condensed Matter and Complex Systems, vol. 38, no. 2, pp. 331-338, (2004).

[8] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. National Acad Sciences, USA, vol. 99, no. 12, pp. 7821-7826, (2002).

[9] G.W. Flake, S. Lawrence, C.L. Giles and F.M. Coetzee, "Self-Organization and identification of Web communities," IEEE Computer, vol. 3, no. 3, pp. 66-71, (2002).

[10] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," Phys. Rev. E, vol. 69, no. 6, pp. 066133-066133, (2004).

[11] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," Phys. Rev. E, vol. 70, no. 6, pp. 264-277, (2004).

[12] A.Clauset, "Finding local community structure in networks," Phys. Rev. E, vol. 72, no. 2Pt2, pp. 254-271, (2005).

[13] A. Lancichinetti, S. Fortunato and J. Kertlsz, "Detecting the overlapping and hierarchical community structure in complex networks," New Journal of Physics, vol. 11, no. 3, pp. 19-44, (2008).

[14] Y Zhou, H Cheng and JX Yu, "Graph Clustering Based on Structural/Attribute Similarities," VLDB'09, ACM, pp. 718-729, (2009).

[15] X Xu, N Yuruk, Z Feng and TAJ Schweiger, "SCAN: a structural clustering algorithm for networks," Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, pp. 824-833, (2007).

[16] Y Lin, T Wang, R Tang, Y Zhou and A Huang, "An Effective Model and Algorithm for Community Detection in Social Networks," Journal of Computer Research and Development, vol. 49, no. 2, pp. 337-345, (2012).

[17] V. Dey, D.K. Pratihar, and G.L. Datta, "Genetic algorithm-tuned entropy-based fuzzy Cmeans algorithm for obtaining distinct and compact clusters," Fuzzy Optimization and Decision Making, vol. 10, no. 2, pp. 153-166, (2011).

[18] Y.Y. Ahn, J.P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," Nature, vol. 466, no. 7307, pp. 761-764, (2010).

[19] L Huang, G Wang, Y Wang, E Blanzieri, and C Su, "Link Clustering with Extended Link Similarity and EQ Evaluation Division," PloS one, vol. 8, no. 6, pp. e66005, 2013).

[20] W.W. Zachary, "An information flow model for conflict and fission in small groups, Journal of Anthropological Research," vol. 33, no. 4, pp. 452-473, (1977).

[21] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behavioral Ecology and Sociobiology, vol. 54, no. 4, pp. 396-405, (2003).