Physica A 405 (2014) 85–91

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient

Yaozu Cui, Xingyuan Wang *, Junqiu Li

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China

HIGHLIGHTS

- Introduce the maximal sub-graphs and the clustering coefficient of two neighboring communities.
- Propose ACC algorithm and an extended modularity.
- Find overlapping vertices.
- Give excellent experimental results.

ARTICLE INFO

Article history:
Received 24 May 2013
Received in revised form 18 February 2014
Available online 12 March 2014

Keywords:
Complex network
Community structures
Maximal sub-graph
Clustering coefficient

ABSTRACT

In this paper, we present an alternate algorithm for detecting overlapping community structures in the complex network. Two concepts named the maximal sub-graph and the clustering coefficient between two neighboring communities are introduced. First, all the maximal sub-graphs are extracted from the original networks and then merge them by considering the clustering coefficient of two neighboring maximal sub-graphs. And a new extended modularity is proposed to quantify this algorithm. The other advantage of this algorithm is that the overlapping vertex can be detected. The effectiveness of our algorithm is tested on some real networks. Finally, we compare the computational complexity of this algorithm with selected close related algorithms. The results show that this algorithm gives satisfactory results.

1. Introduction

Many real-world systems in nature and society can be described as complex networks or graphs [1–3]. The entities of the system are represented by the nodes (vertices) and the interactions between the entities are represented by the edges. Examples include social relationships, spreading of viruses and diseases, and the Internet and the World Wide Web [4–9], etc. The research of complex networks is important to understand the structure of the networks and the interaction of the entities in the networks. In the past decade, a lot of features about the complex networks could be obtained from studying the complex networks. One common feature is the community structure [10,11], where the nodes within a community are higher connected to each other than the nodes among communities.

The detection of community structures has attracted much attention from various real networks. Many methods have been proposed to detect the community structures of complex networks and applied successfully to some real complex networks [12–22]. On the basis of the process, results, and other features, these methods can be roughly classified into two categories. One category is the graph partitioning method, where each vertex belongs to only one community. Different from

[*] Corresponding author.
E-mail addresses: cyz3471@sina.com (Y. Cui), wangxy@dlut.edu.cn (X. Wang), meiliqiutian@126.com (J. Li).

http://dx.doi.org/10.1016/j.physa.2014.03.027
0378-4371/© 2014 Elsevier B.V. All rights reserved.

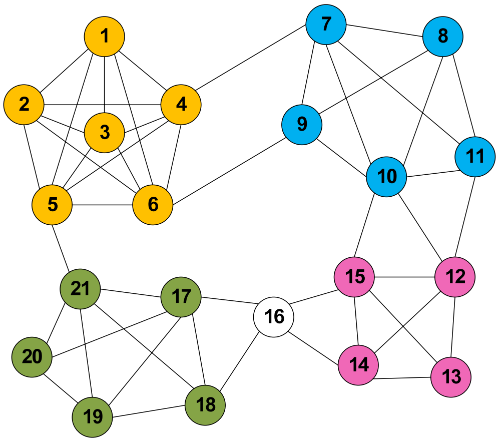

Fig. 1. A schematic network represented by four different colors.

classical graph partitioning problem, for example, Kernighan–Lin algorithm [23] and the spectral bisection algorithm [24,25], the number of communities and the size of each community are known before detecting the communities in the complex networks. Unfortunately, there is no commonly accepted standard to evaluate the detection of communities.

The other category is the hierarchical clustering method based on discovering the high connected vertices. And this category can be divided into the agglomerative method and divisive method by adding or removing edges [26]. In this kind, some vertices may belong to more than one community, so overlapping structure can exist. In 2002, Newman and Girvan [10] proposed the modularity (represented by a Q function) to evaluate the goodness of the partition. Then some detecting community methods have been proposed based on the optimizing modularity [27–29]. This kind of methods is useful to understand the community structures of networks, especially for the small networks. However, the optimization of the modularity has a resolution limit problem [30,31]. And the methods based on the modularity cannot detect overlapping vertices.

In this paper, ACC algorithm (based on the Clustering Coefficient of two neighboring maximal sub-graphs) is proposed to detect the overlapping community structures in complex networks. And an extended modularity based on the clustering coefficient between two neighboring maximal sub-graphs is proposed to quantify ACC algorithm. The first step of this algorithm is to extract all the maximal sub-graphs from a given complex network. Then calculate the clustering coefficient of two neighboring maximal sub-graphs to obtain the community structures. The other advantage of this algorithm is that the overlapping vertices can be found accurately. In Section 2 we explain the maximal sub-graphs and propose ACC algorithm in detail. We give a number of tests of this algorithm on some real-world networks in Section 3. In Section 4, the conclusions are given.

2. The algorithm

Actually overlapping and hierarchical exist at the same time in the some real networks. Overlapping means that some vertices do not belong to only one community, that is to say, they may belong to more than one community. Hierarchical means that communities can be detected into some smaller communities or some communities may be merged into a larger community. In this paper, we devote to detection overlapping and hierarchical community structures.

2.1. The maximal sub-graphs

As to the network in Fig. 1, the overlapping community structures can be distinctly represented by four different colors, that are {1, 2, 3, 4, 5, 6}, {7, 8, 9, 10, 11, 12}, {12, 13, 14, 15, 16} and {16, 17, 18, 19, 20, 21}. The vertex {1, 2, 3, 4, 5, 6} consists of the highest connective community than the rest in the complex network. We call such a complete sub graph as the maximal sub-graph., which is not a subset of any other communities in a complex network. In Table.1, we employ the algorithm proposed in Ref. [32] to extract all the maximal sub-graphs from Fig. 1, which is an important step during detecting the community structures in the complex networks.

2.2. The clustering g coefficient of two neighboring communities

The traditional definition of clustering coefficient has global algorithm and local algorithm. The global algorithms base on computing the density of triplet, but the local algorithms concern to compute the density of vertices. In a network, vertex i has k_i edges directly connecting to other k_i vertices, that is to say, k_i vertices are the neighboring vertices of vertex i. These are at most $k_i(k_i - 1)/2$ edges among these k_i vertices. Let E_i be the number of the actual edges existing among these k_i

Table 1
All the maximal sub-graphs of Fig. 1 with 21 vertices.

The label	All the maximal sub-graphs
a	{1, 2, 3, 4, 5, 6}
b	{7, 8, 9, 10}
c	{7, 8, 10, 11}
d	{12, 13, 14, 15}
e	{14, 15, 16}
f	{16, 17, 18}
g	{17, 18, 19, 21}
h	{17, 19, 20, 21}

Table 2
The clustering coefficient of two neighboring maximal sub-graphs [C_c] in the network with 21 vertices.

	a	b	c	d	e	f	g	h
a	1	0.511	0.489	0	0	0	0.489	0.489
b	0.511	1	0.9	0.5	0.476	0	0	0
c	0.489	0.9	1	0.5	0.476	0	0	0
d	0	0.5	0.5	1	0.8	0.534	0	0
e	0	0.476	0.476	0.8	1	0.6	0.534	0.476
f	0	0	0	0.534	0.6	1	0.8	0.733
g	0.489	0	0	0	0.534	0.8	1	0.9
h	0.489	0	0	0	0.476	0.733	0.9	1

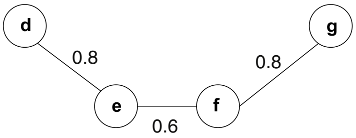

Fig. 2. The clustering coefficient of four neighboring maximal sub-graphs.

vertices. Then, the ratio of the existing and the possible numbers of edges among k_i nodes is defined to be the clustering coefficient of vertex i, denoted C_i [33]; namely,

$$C_i = \frac{2E_i}{k_i(k_i - 1)}. \tag{1}$$

However, in this paper, the clustering coefficient of two neighboring communities [C_c] is proposed to whether to merge two neighboring communities into a new larger community. If the total number of two communities is n_c, these n_c vertices will have $n_c(n_c - 1)/2$ edges as a complete graph and they may have F_c edges in fact. Then, the clustering coefficient of two neighboring communities [C_c] is defined as follows

$$C_c = \frac{2F_c}{n_c(n_c - 1)}. \tag{2}$$

After all the maximal sub-graphs are extracted, each maximal sub-graph is as an independent community. Then we will calculate the clustering coefficient of two neighboring communities [C_c] whether or not be combined into a new larger community. For clarification, consider the network of 21 vertices in Fig. 1. First, 8 maximal sub-graphs are extracted in Table 1 and each one will be given a new label. We calculate the clustering coefficient of two neighboring communities [C_c] in Table 2. We introduce the symmetric matrices actually in Table 2. The maximum value of the C_c in Table 2 is $C_{c\,\max} = 1$, which is the clustering coefficient of the maximal sub-community and itself.

However, it can be seen that there is no constraint to have priority to combine two communities into a larger community. To overcome this, we can consider a threshold value δ for the clustering coefficient of two neighboring communities [C_c]. For $0.476 \leq \delta \leq 0.534$, the two neighboring maximal sub-graphs cannot be merged into a new larger community. On the contrary, the two neighboring maximal sub-graphs can be merged into a new larger community for $0.6 \leq \delta \leq 0.9$.

There is a kind of special cases. In Fig. 2, e can be merged with d and f, and the clustering coefficient between e and $d$is higher than the clustering coefficient between e and f, so e and d will be preferentially merged into a new community. In the same way, f and g will be preferentially merged into a new community. And the common vertices between e and f are the overlapping vertices. So vertex 6 is the overlapping vertex in the small network with 21 vertices. In short, for $0.6 \leq \delta < 1$, the two maximal sub-graphs can be merged into a new larger community. For $\delta < 0.6$, all the maximal sub-graphs cannot be merged into one cluster. Here there is only one community, i.e. the network itself.

2.3. Extended modularity

As well known, the modularity defined by Newman and Girvan [13] was used to measure the quality of detection community structures in the unweighted networks. But an extended modularity was proposed by Newman to measure the goodness of a partition of weighted networks. Nicosia et al. [34] proposed another modularity to evaluate the quality of overlapped community structures. In this paper, given an unweighted and undirected network, the modularity based the clustering coefficient of all the maximal sub-graphs [C_c] can be formalized as

$$Q_C = \frac{1}{2m}\sum_{ij}\left[A_{ij}^{G} - \frac{E_{G_i}E_{G_j}}{2m}\right]C_c(G_i, G_j), \tag{3}$$

where A_{ij}^{G} represents an adjacency matrix with the clustering coefficient of all the maximal sub-graphs between i and j, E_{G_i} is the number of edges in a maximal sub-graph i, m is the total number of edges in the network. $C_c\left(G_i, G_j\right)$ is the clustering coefficient of two maximal sub-graphs.

2.4. ACC algorithm

Our ACC algorithm based on the clustering coefficient of two neighboring maximal sub-graphs can be summarized as follows:

(a) initially, use the algorithm [32] to extract all the maximal sub-graphs from given complex networks with the new label;
(b) then, calculate the number of edges in each maximal sub-graph;
(c) find all neighbors of each maximal sub-graph, and calculate the clustering coefficient [C_c];
(d) consider a threshold value δ whether to combine two communities into a new community and to determine the overlapping vertices. Then calculate the modularity [Q_C];
(e) repeat steps (c) and (d);
(f) quit the repetition and get a proper community structure.

Analyzing the computational complexity of ACC algorithm, the steps of extracting all the maximal sub-groups has computational complexity of at most $O(n)$. When the number of the maximal sub-groups is $x(x \leq n)$, take $O(x)$ to calculate the number of edges in each maximal sub-graph and take $O(x^2)$ to calculate the clustering coefficient [C_c]. Finally, the worst computing time is upper bound at most $O(n^2)$.

3. Applications and results

In this section, we implement our overlapping community detecting algorithm ACC by Java programming language running on a PC with 2.67 GHz processor, 4 GB memory and Windows7 operating system. And extensively test it on some real networks with known community structures.

First, a real network is Zachary's karate club network [35], which is widely used as a benchmark for testing the methods of detection community structures, which is an un-weighted network with 34 members of a karate club as vertices and 78 edges representing friendships among members of the club. However, a dispute developed between the club's administrator and its principal karate teacher, which results the club eventually splits into two smaller clubs, centered at the administrator and the teacher respectively. The communities of this network are depicted in Fig. 3. The nodes 1 and 34 represented the administrator and the teacher respectively. Similar to many existing community detection algorithms, our ACC algorithm detects the network into two overlapping communities represented by two circle and square. This partition corresponds to the modularity [Q_C] with the value 0.357. For $0.600 \leq \delta < 1$, there are two communities and two overlapping vertices; for $0.733 \leq \delta < 1$, there are three communities which are labeled by different colors in Fig. 3 and the modularity [Q_C] is 0.408. With the increasing the number of community, the modularity [Q_C] continues to increase. The overlapping vertices consist of 3 and 10, shared by two communities. Except these overlapping vertices, the result reflects the real split of this network.

Second, we also test our ACC algorithm on another real data set taken from [36], which describes the associations between 62 dolphins living in Doubtful Sound, New Zealand. For $0.627 \leq \delta < 1$, two communities are naturally detected by our ACC algorithm in Fig. 4, represented by squares and circles. This partition corresponds to the modularity [Q_C] with the value 0.485. And vertex 40 is the overlapping vertex. For $0.833 \leq \delta < 1$, we detect three communities and two overlapping vertices, represented by different colors as shown in Fig. 3. For $\delta < 0.6$, all the maximal sub-graphs cannot combine into a larger community and only one community is detected, the original network itself. Then we can see the hierarchical community structures in these networks by ACC algorithm.

Then, we have also used normalized mutual information [37] to measure the detection algorithm. In Fig. 5 we plot the performance of three algorithms testing a set of computer-generated networks presented by Girvan and Newman [10]. As we can see from the figure, from $k_{out} = 0$ to $k_{out} = 6$, the natural communities are detected by both the three algorithms. When $k_{out} > 6$, the accuracy of this three algorithm begin to drop down markedly. However, ACC algorithm does perform well than GN algorithm and EAGLE.

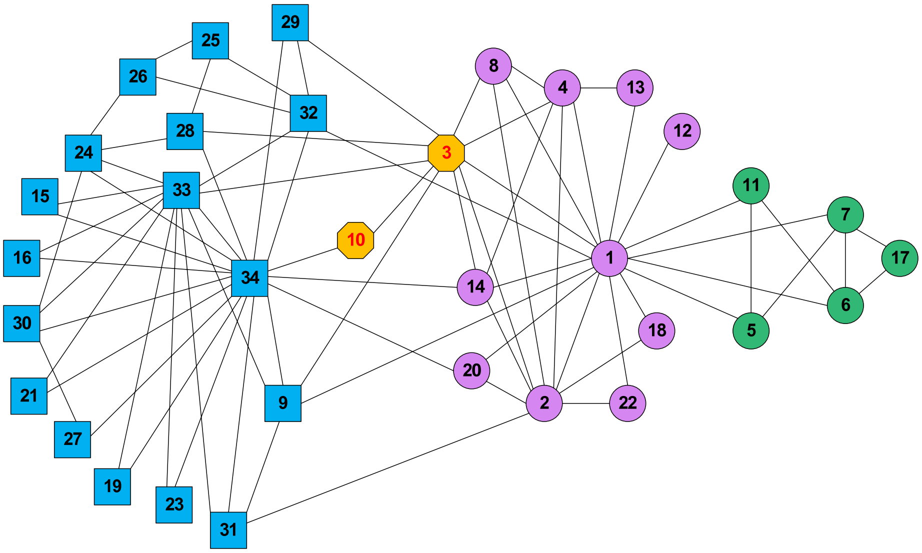

Fig. 3. The overlapping communities from Zachary's karate club study.

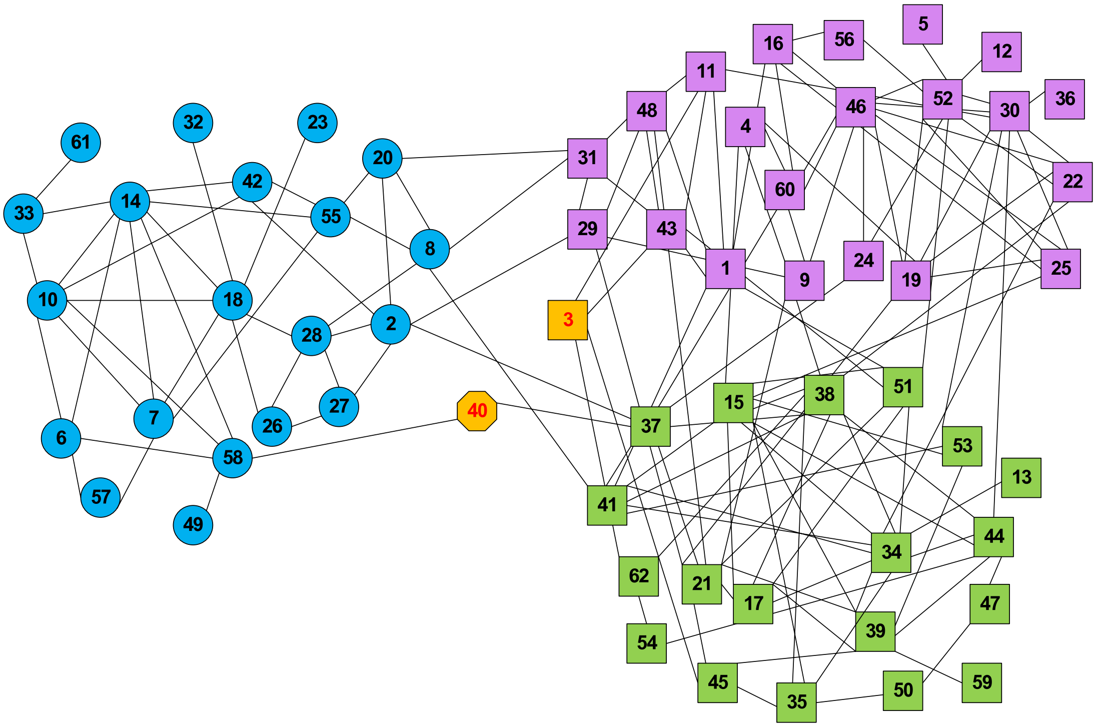

Fig. 4. The community structures detected by ACC algorithm from Lusseau's network of dolphins.

As shown in Fig. 6, we compare ACC algorithm with four recently reported algorithms (GN [13], BGLL [32], Chen [21], EAGLE [20]) testing some real world networks. That is, the friendship network from Zachary's karate club study with 34 vertices and 78 edges [33], college football with 115 nodes and 615 edges [10], netscience network with 1589 vertices and 2742 edges [38], dolphin's associations with 62 nodes and 159 edges [36], cond-mat-2003 with 31163 nodes and 120029 edges [39]. These networks can be downloaded from http://www-personal.umich.edu/~mejn/netdata/. The execution time of our proposed algorithm is low in comparison with the other four algorithms.

4. Conclusions

We proposed an algorithm for detecting the community structures in the complex networks. In this algorithm, all the maximal sub-graphs first extracted and then merged with neighboring maximal sub-graphs to obtain the communities

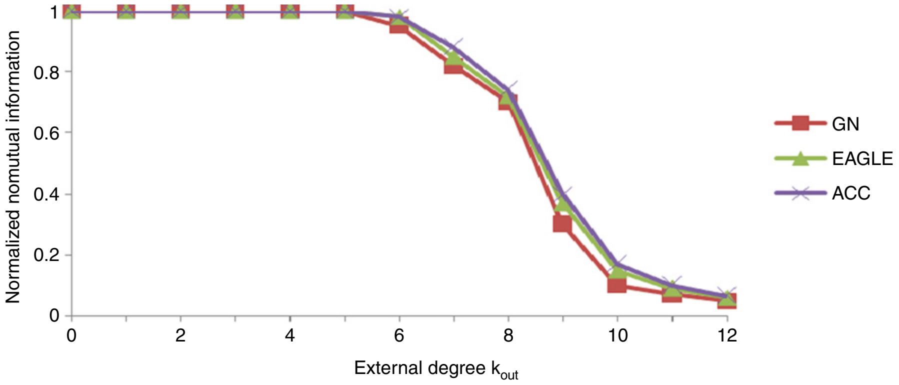

Fig. 5. The test on the benchmark of Girvan and Newman.

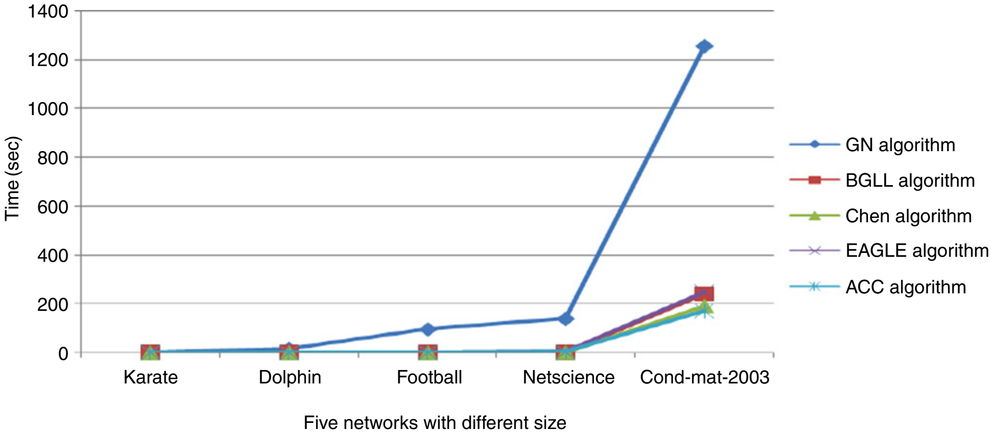

Fig. 6. The execution time of our proposed ACC algorithm in comparison with other four existing algorithms.

using clustering coefficient of two neighboring sub-graphs. An extended modularity is proposed to quantify our algorithm. We have tested our algorithm on some real networks. The results indicate that the overlapping vertices and hierarchical community structures are detected. The computation cost is verified in Fig. 6. All this make it a useful and efficient algorithm for detecting overlapping community structures.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Nos: 61370145, 61173183, and 60973152), the Doctoral Program Foundation of Institution of Higher Education of China (No: 20070141014), Program for Liaoning Excellent Talents in University (No: LR2012003), the National Natural Science Foundation of Liaoning province (No: 20082165) and the Fundamental Research Funds for the Central Universities (No: DUT12JB06).

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97.
[2] S.H. Strogatz, Exploring complex networks, Nature 410 (6825) (2001) 268–276.
[3] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[4] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, A. Arenas, Models of social networks based on social distance attachment, Phys. Rev. E 70 (5) (2004) 056122.
[5] Z.W. Liu, H.G. Zhang, Q.L. Zhang, Novel stability analysis for recurrent neural networks with multiple delays via line integral-type L–K functional, IEEE Trans. Neural Netw. 21 (11) (2010) 1710–1718.
[6] F. Képès, Biological Networks, World Scientific, Singapore, 2007.
[7] A. Vazquez, R. Pastor-Satorras, A. Vespignani, Large-scale topological and dynamical properties of the Internet, Phys. Rev. E 65 (6) (2002) 066130.
[8] Y.J. Liu, C.L.P. Chen, G.X. Wen, S.C. Tong, Adaptive neural output feedback tracking control for a class of uncertain iscrete-time nonlinear systems, IEEE Trans. Neural Netw. 22 (7) (2011) 1162–1167.
[9] K.A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen, Modularity and extreme edges of the internet, Phys. Rev. Lett. 90 (14) (2003) 148701.
[10] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (6) (2002) 821–7824.
[11] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.
[12] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (6) (2004) 066133.

[13] H.G. Zhang, T.D. Ma, G.B. Huang, C.X. Wang, Robust global exponential synchronization of uncertain chaotic delayed neural networks via dual-stage impulsive control, IEEE Trans. Syst. Man Cybern. B 40 (3) (2010) 831–844.
[14] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[15] F. Wei, W.N. Qian, C. Wang, A.Y. Zhou, Detecting overlapping community structures in networks, World Wide Web 12 (2) (2009) 235–261.
[16] Y.J. Liu, Y.Q. Zheng, Adaptive robust fuzzy control for a class of uncertain chaotic systems, Nonlinear Dynam. 57 (3) (2009) 431–439.
[17] X. Liu, J.Y.- L. Forrest, Q. Luo, D.Y. Yi, Detecting community structure using biased random merging, Physica A 391 (4) (2012) 1797–1810.
[18] A. Lancichinetti, S. Fortunato, J. Kertèsz, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (3) (2009) 033015.
[19] L.L. Cui, H.G. Zhang, B. Chen, Q.L. Zhang, Asymptotic tracking control scheme for mechanical systems with external disturbances and friction, Neurocomputing 73 (7–9) (2010) 1293–1302.
[20] H.W. Shen, X.Q. Cheng, K. Cai, M.B. Hu, Dectect overlapping and hierarchical community structure in networks, Physica A 388 (8) (2009) 1706–1712.
[21] D.B. Chen, Y. Yu, M.S. Shang, A fast and efficient heuristic algorithm for detecting community structures in complex networks, Physica A 388 (13) (2009) 2741–2749.
[22] Y.J. Liu, S.C. Tong, D. Wang, T.S. Li, C.L.P. Chen, Adaptive neural output feedback controller design with reduced-order observer for a class of uncertain nonlinear SISO systems, IEEE Trans. Neural Netw. 22 (8) (2011) 1328–1334.
[23] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, Bell Syst. Tech. J. 49 (2) (1970) 291–307.
[24] M. Fiedler, Algebraic connectivity of graphs, Czechoslovak Math. J. 23 (98) (1973) 298–305.
[25] A. Pothen, H.D. Simon, K.P. Liou, Partitioning sparse matrices with eigenvectors of graphs, SIAM. J. Matrix Anal. Appl. 11 (3) (1990) 430–452.
[26] J. Scott, Social Network Analysis: A Handbook, second ed., Sage Publications, London, 2002.
[27] Z.H. Wu, Y.F. Lin, H.Y. Wan, S.F. Tian, K.Y. Hu, Efficient overlapping community detection in huge real-world networks, Physica A 391 (7) (2012) 2475–2490.
[28] T. Nepusz, A. Petróczi, L. Négyessy, F. Bazsó, Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77 (1) (2008) 016107.
[29] Y.P. Li, Y.M. Ye, E.K. Wang, Fast computation of modularity in agglomerative clustering methods for community discovery, Int. J. Adv. Comput. Technol. 3 (4) (2011) 153–164.
[30] S. Fortunato, M. Barthélemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. USA 104 (1) (2007) 36–46.
[31] J.M. Kumpula, J. Saramaki, K. Kaski, J. Kertesz, Limited resolution limit in complex network community detection with Potts model approach, Eur. Phys. J. B 56 (1) (2007) 41–45.
[32] V.D. Blondel, J.L. Guillaume, R. Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. 2008 (10) (2008) 10008.
[33] G.R. Chen, X.F. Wang, X. Li, Introduction to Complex Networks: Models, Structures and Dynamics, Higher Education Press, Beijing, 2012.
[34] V. Nicosia, G. Mangioni, V. Carchiolo, M. Malgeri, Extending the definition of modularity to directed graphs with overlapping communities, J. Stat. Mech. 2009 (3) (2009) 03024.
[35] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473.
[36] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol 54 (2003) 396–405.
[37] L. Danon, A. Díaz-Guilera, A. Arenas, Comparing community structure identification, J. Stat. Mech. 2005 (09) (2005) 09008.
[38] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.
[39] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Nat. Acad. Sci. USA 98 (2) (2001) 404–409.