Available online at www.sciencedirect.com
ScienceDirect
Physica A 374 (2007) 483–490
PHYSICA A
www.elsevier.com/locate/physa

Identification of overlapping community structure in complex networks using fuzzy *c*-means clustering

Shihua Zhang[a,*], Rui-Sheng Wang[b], Xiang-Sun Zhang[a]

[a]*Academy of Mathematics & Systems Science, Chinese Academy of Science, Beijing 100080, China*
[b]*School of Information, Renmin University of China, Beijing 100872, China*

Received 28 June 2006
Available online 7 August 2006

Abstract

Identification of (overlapping) communities/clusters in a complex network is a general problem in data mining of network data sets. In this paper, we devise a novel algorithm to identify overlapping communities in complex networks by the combination of a new modularity function based on generalizing NG's Q function, an approximation mapping of network nodes into Euclidean space and fuzzy *c*-means clustering. Experimental results indicate that the new algorithm is efficient at detecting both good clusterings and the appropriate number of clusters.

Keywords: Overlapping community structure; Modular function; Spectral mapping; Fuzzy *c*-means clustering; Complex network

1. Introduction

Large complex networks representing relationships among set of entities have been one of the focuses of interest of scientists in many fields in the recent years. Various complex network examples include social network, worldwide web network, telecommunication network and biological network. One of the key problems in the field is 'How to describe/explain its community structure'. Generally, a community in a network is a subgraph whose nodes are densely connected within itself but sparsely connected with the rest of the network. Many studies have verified the community/modularity structure of various complex networks such as protein-protein interaction network, worldwide web network and co-author network. Clearly, the ability to detect community structure in a network has important practical applications and can help us understand the network system.

Although the notion of community structure is straightforward, construction of an efficient algorithm for identification of the community structure in a complex network is highly nontrivial. A number of algorithms for detecting the communities have been developed in various fields (for a recent review see Ref. [1] and a recent comparison paper see Ref. [2]). There are two main difficulties in detecting community structure. The first is that we don't know how many communities there are in a given network. The usual drawback in many

*Corresponding author.
E-mail addresses: zsh@amss.ac.cn (S. Zhang), wrs@ruc.edu.cn (R.-S. Wang), zxs@amt.ac.cn (X.-S. Zhang).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2006.07.023

algorithms is that they cannot give a valid criterion for measuring the community structure. Secondly, it is a common case that some nodes in a network can belong to more than one community. This means the overlapping community structure in complex networks. Overlapping nodes may play a special role in a complex network system. Most known algorithms such as divisive algorithm [3–5] cannot detect them. Only a few community-detecting methods [6,7] can uncover the overlapping community structure. Taking into account the first difficulty, Newman and Girvan [8] has developed a new approach. They introduced a modularity function Q for measuring community structure. In order to write the context properly, we refer to a similar formulation in Ref. [5]. In detail, given an undirected graph/network $G(V, E, W)$ consisting of the node set V, the edge set E and a symmetric weight matrix $W = [w_{ij}]_{n\times n}$, where $w_{ij} \geqslant 0$ and n is the size of the network, the modularity function Q is defined as

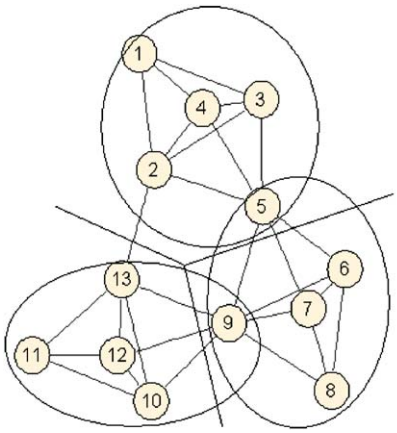
$$Q(P_k) = \sum_{c=1}^{k} \left[\frac{L(V_c, V_c)}{L(V, V)} - \left(\frac{L(V_c, V)}{L(V, V)} \right)^2 \right], \tag{1}$$

where P_k is a partition of the nodes into k groups and $L(V', V'') = \sum_{i\in V', j\in V''} w(i,j)$. The Q function measures the quality of a given community structure organization of a network and can be used to automatically select the optimal number of communities k according to the maximum Q value [8,5]. The measure has been used for developing new detection algorithms such as Refs. [5,9,4]. White and Smyth [5] showed that optimizing the Q function can be reformulated as a spectral relaxation problem and proposed two spectral clustering algorithms that seek to maximize Q.

In this study, we develop an algorithm for detecting overlapping community structure. The algorithm combines the idea of modularity function Q [8], spectral relaxation [5] and fuzzy c-means clustering method [10] which is inspired by the general concept of fuzzy geometric clustering. The fuzzy clustering methods don't employ hard assignment, while only assign a membership degree u_{ij} to every node v_i with respect to the cluster C_j.

2. Method

Simulation across a wide variety of simulated and real world networks showed that large Q values are correlated with better network clusterings [8]. Then maximizing the Q function can obtain final 'optimal' community structure. It is noted that in many complex networks, some nodes may belong to more than one community. The divisive algorithms based on maximizing the Q function fail to detect such case. Fig. 1 shows an example of a simple network which visually suggests three clusters and classifying node 5 (or node 9) into



Method: K-means

k = 1 2
Q = 0 0.181
3 4 5
0.223 0.158 -0.032

Fuzzy c-means
lembda=0.15
k = 1 2 3 4 5
Q = 0 0.181 0.326 0.168 0.006

clusters	Soft assignment	Final clusters
1 [1 0 0]	[0.9951 0.0026 0.0023]	[1 0 0]
2 [1 0 0]	[0.9804 0.0108 0.0088]	[1 0 0]
3 [1 0 0]	[0.9984 0.0008 0.0008]	[1 0 0]
4 [1 0 0]	[0.9984 0.0008 0.0008]	[1 0 0]
5 [1 0 0]	[0.4327 0.1133 0.4540]	[1 0 1]
6 [0 0 1]	[0.0039 0.0037 0.9924]	[0 0 1]
7 [0 0 1]	[0.0039 0.0037 0.9924]	[0 0 1]
8 [0 0 1]	[0.0012 0.0013 0.9975]	[0 0 1]
9 [0 0 1]	[0.0715 0.1519 0.7766]	[0 1 1]
10 [0 1 0]	[0.0020 0.9959 0.0022]	[0 1 0]
11 [0 1 0]	[0.0012 0.9976 0.0011]	[0 1 0]
12 [0 1 0]	[0.0020 0.9959 0.0022]	[0 1 0]
13 [0 1 0]	[0.0054 0.9899 0.0047]	[0 1 0]

Fig. 1. An example of network showing Q and $\widetilde{Q}$ values for different number k of clusters using the same spectral mapping but different cluster methods, i.e. k-means and fuzzy c-means, respectively. For the latter, it shows every node's soft assignment and membership of final clusters with $\lambda = 0.15$.

two clusters at the same time may be more appropriate intuitively. So we introduce the concept of fuzzy membership degree to the network clustering problem in the following subsection.

2.1. A new modular function

If there are k communities in total, we define a corresponding $n \times k$ 'soft assignment' matrix $U_k = [u_1, \ldots, u_k]$ with $0 \leqslant u_{ic} \leqslant 1$ for each $c = 1, \ldots, k$ and $\sum_{c=1}^{k} u_{ic} = 1$ for each $i = 1, \ldots, n$. With this we define the membership of each community as $\bar{V}_c = \{i | u_{ic} > \lambda, i \in V\}$, where λ is a threshold that can convert a soft assignment into final clustering. We define a new modularity function $\widetilde{Q}$ as

$$\widetilde{Q}(U_k) = \sum_{c=1}^{k} \left[\frac{A(\bar{V}_c, \bar{V}_c)}{A(V, V)} - \left(\frac{A(\bar{V}_c, V)}{A(V, V)} \right)^2 \right], \tag{2}$$

where U_k is a fuzzy partition of the vertices into k groups and $A(\bar{V}_c, \bar{V}_c) = \sum_{i \in \bar{V}_c, j \in \bar{V}_c} ((u_{ic} + u_{jc})/2) w(i,j)$, $A(\bar{V}_c, V) = A(\bar{V}_c, \bar{V}_c) + \sum_{i \in \bar{V}_c, j \in V \setminus \bar{V}_c} ((u_{ic} + (1 - u_{jc}))/2) w(i,j)$ and $A(V, V) = \sum_{i \in V, j \in V} w(i,j)$. This of course can be thought as a generalization of the Newman's Q function. Our objective is to compute a soft assignment matrix by maximizing the new Q function with appropriate k. How could we do?

2.2. Spectral mapping

White and Smyth [5] showed that the problem of maximizing the modularity function Q can be reformulated as an eigenvector problem and devised two spectral clustering algorithms. Their algorithms are similar in spirit to a class of spectral clustering methods which map data points into Euclidean space by eigendecomposing a related matrix and then grouping them by general clustering methods such as k-means and hierarchical clustering [5,9]. Given a network and its adjacent matrix $A = (a_{ij})_{n \times n}$ and a diagonal matrix $D = (d_{ii})$, $d_{ii} = \sum_k a_{ik}$, two matrices $D^{-1/2} A D^{-1/2}$ and $D^{-1} A$ are often used. A recent modification [11] uses the top K eigenvectors of the generalized eigensystem $Ax = tDx$ instead of the K eigenvectors of the two matrices mentioned above to form a matrix whose rows correspond to original data points. The authors show that after normalizing the rows using Euclidean norm, their eigenvectors are mathematically identical and emphasize that this is a numerically more stable method. Although their result is designed to cluster real-valued points [11,12], it is also appropriate for network clustering. So in this study, we compute the top $k - 1$ eigenvectors of the eigensystem to form a $(k - 1)$-dimensional embedding of the graph into Euclidean space and use 'soft-assignment' geometric clustering on this embedding to generate a clustering U_k (k is the expected number of clusters).

2.3. Fuzzy c-means

Here, in order to realize our 'soft assignment', we introduce fuzzy c-means (FCM) clustering method [10,13] to cluster these points and maximize the $\widetilde{Q}$ function. Fuzzy c-means is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 [10] and improved by Bezdek in 1981 [13]) is frequently used in pattern recognition. It is based on minimization of the following objective function

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \|x_i - c_j\|^2, \tag{3}$$

over variables u_{ij} and c with $\sum_j u_{ij} = 1$. $m \in [1, \infty)$ is a weight exponent controlling the degree of fuzzification. u_{ij} is the membership degree of x_i in the cluster j. x_i is the ith d-dimensional measured data point. c_j is the d-dimensional center of the cluster j, and $\| * \|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership degree u_{ij} and the cluster centers c_j. This procedure converges to a local minimum or a saddle point of J_m.

2.4. The flow of the algorithm

Given an upper bound K of the number of clusters and the adjacent matrix $A = (a_{ij})_{n\times n}$ of a network. The detailed algorithm is stated straightforward for a given λ as follows:

- Spectral mapping:
  (i) Compute the diagonal matrix $D = (d_{ii})$, where $d_{ii} = \sum_k a_{ik}$.
  (ii) Form the eigenvector matrix $E_K = [e_1, e_2, \ldots, e_K]$ by computing the top K eigenvectors of the generalized eigensystem $Ax = tDx$.
- Fuzzy c-means: for each value of k, $2 \leqslant k \leqslant K$:
  (i) Form the matrix $E_k = [e_2, e_3, \ldots, e_k]$ from the matrix E_K.
  (ii) Normalize the rows of E_k to unit length using Euclidean distance norm.
  (iii) Cluster the row vectors of E_k using fuzzy c-means or any other fuzzy clustering method to obtain a soft assignment matrix U_k.
- Maximizing the modular function: Pick the k and the corresponding fuzzy partition U_k that maximizes $\widetilde{Q}(U_k)$.

In the algorithm above, we initialize FCM such that the starting centroids are chosen to be as orthogonal as possible which is suggested for k-means clustering method in Ref. [12]. The initialization does not change the time complexity, and also can improve the quality of the clusterings, thus at the same time reduces the need for restarting the random initialization process.

The framework of our algorithm is similar to several spectral clustering methods in previous studies [5,9,12,11]. We also map data points (i.e. network nodes in our study) into Euclidean space by computing the top K eigenvectors of a generalized eigen system and then cluster the embedding using a fuzzy clustering method just as others using geometric clustering algorithm or general hierarchical clustering algorithm. Here, we emphasize two key points different from those earlier studies:

- We introduce a generalized modular function $\widetilde{Q}$ employing fuzzy concept, which is devised for evaluating the goodness of overlapping community structure.
- In combination with the novel $\widetilde{Q}$ function, we introduce fuzzy clustering method into network clustering instead of general hard clustering methods.

This means that our algorithm can uncover overlapping clusters, whereas general framework: "Objective function such as Q function and Normalized cut function + Spectral mapping + general geometric clustering/hierarchical clustering" cannot achieve this.

3. Experimental results

We have implemented the proposed algorithm by Matlab. And the fuzzy clustering toolbox [14] is used for our analysis. In order to make an intuitive comparison, we also compute the hard clustering based on the original Q-function, spectral mapping (same as we used) and k-means clustering. We illustrate the fuzzy concept and the difference of our method with traditional divisive algorithms by a simple example shown in Fig. 1. Just as mentioned above, the network visually suggests three clusters. But classifying node 5 (or node 9) simultaneously into two clusters may be more reasonable. We can see from Fig. 1 that our method did uncover the overlapping communities for this simple network, while the traditional method can only make one node belong to a single cluster.

We also present the analysis of two real networks, i.e. the Zachary's karate club network and the American college football team network for better understanding the differences between our method and traditional methods.

3.1. Zachary's karate club

The famous karate club network analyzed by Zachary [15] is widely used as a test example for methods of detecting communities in complex networks [1,8,16,3,4,17,9,18,19]. The network consists of 34 members of a karate club as nodes and 78 edges representing friendship between members of the club which was observed over a period of two years. Due to a disagreement between the club's administrator and the club's instructor, the club split into two smaller ones. The question we concern is that if we can uncover the potential behavior of the network, detect the two communities or multiple groups, and particularly identify which community a node belongs to. The network is presented in Fig. 2, where the squares and the circles label the members of the two groups. The results of k-means and our analysis are illustrated in Fig. 3.

The k-means combined with Q function divides the network into three parts (see in Fig. 3A), but we can see that some nodes in one cluster are also connected densely with another cluster such as node 9 and 31 in cluster 1 densely connecting with cluster 2, and node 1 in cluster 2 with cluster 3. Fig. 3B shows the results of our method, from which we can see that node 1, 9, 10, 31 belong to two clusters at the same time. These nodes in the network link evenly with two clusters. Another thing is that the two methods both uncover three communities but not two. There is a small community included in the instructor's faction, since the set of nodes 5, 6, 7, 11, 17 only connects with node 1 in the instructor's faction. Note that our method also classifies node 1 into the small community, while k-means does not.

3.2. Network of American college football teams

The second network we have investigated is the college football network which represents the game schedule of the 2000 season of Division I of the US college football league. The nodes in the network represent the 115 teams, while the links represent 613 games played in the course of the year. The teams are divided into conferences of 8–12 teams each and generally games are more frequent between members of the same conference than between teams of different conferences. The natural community structure in the network makes it a commonly used workbench for community-detecting algorithm testing [3,5,7].

Fig. 4 shows how the modularity Q and $\widetilde{Q}$ vary with k with respect to k-means and our method, respectively. The peak for k-means is at $k = 12$, $Q = 0.5398$, while for our algorithm at $k = 10$, $\widetilde{Q} = 0.4673$ with $\lambda = 0.10$. Both methods identify ten communities which contain ten conferences almost exactly. Only teams labeled as Sunbelt are not recognized as belonging to a same community for both methods. This group is classified as well in the results of Refs. [3,19]. This happens because the Sunbelt teams played nearly as many games against Western Athletic teams as they played in their own conference, and they also played quite a number of games against Mid-American team. Our method identified 11 nodes (teams) which belong to at least two

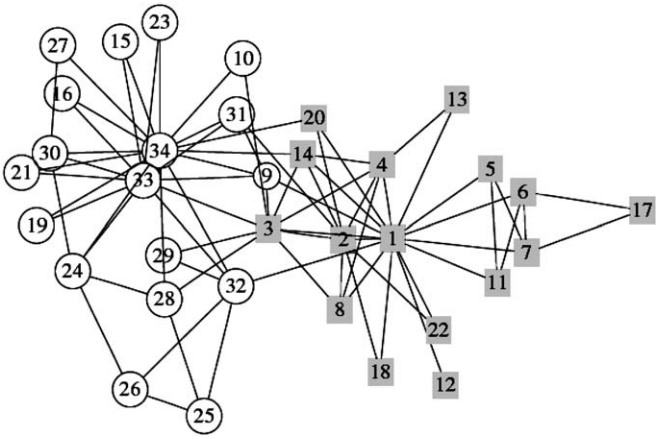


Fig. 2. Zachary's karate club network. Square nodes and circle nodes represent the instructor's faction and the administrator's faction, respectively. This figure is from Newman and Girvan [8].

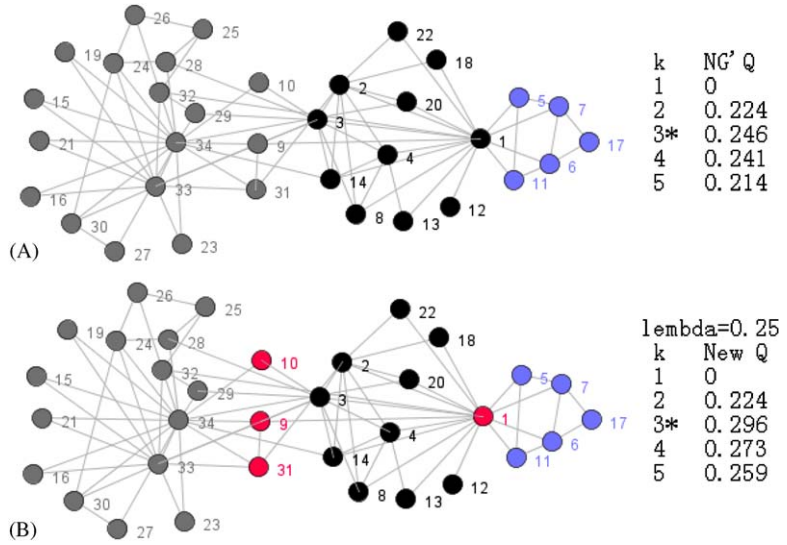


Fig. 3. The results of both k-means and our method applied to karate club network. A: The different colors represent three different communities obtained by k-means and the right table shows values of NG'Q versus different k. B: Four red nodes represent the overlap of two adjacent communities obtained by our method and the right table shows values of new Q versus different k with $\lambda = 0.25$.

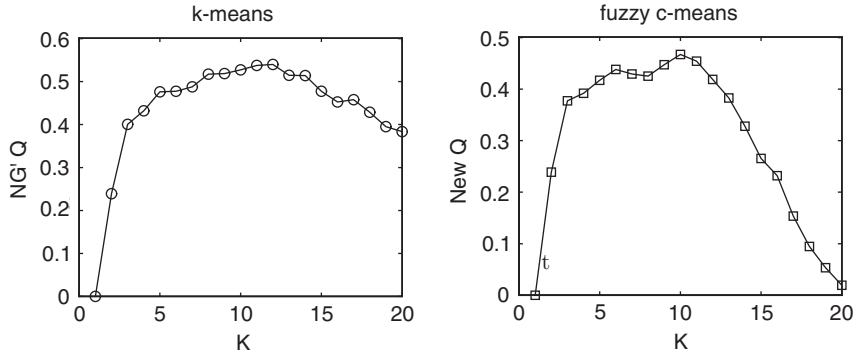


Fig. 4. Q and $\widetilde{Q}$ values versus k with respect to k-means and fuzzy c-means clustering methods for the network of American college football team.

communities (see Fig. 5, 11 red nodes). These nodes generally connect evenly with more than one community, so we cannot classify them into one specific community correctly. These nodes represent 'fuzzy' points which cannot be classified correctly by employing current link information. Maybe such points play a 'bridge' role in two or more communities in complex network of other types.

4. Conclusion and discussion

In this paper, we present a new method to identify the community structure in complex networks with a fuzzy concept. The method combines a generalized modularity function, spectral mapping, and fuzzy clustering technique. The nodes of the network are projected into d-dimensional Euclidean space which is obtained by computing the top d nontrivial eigenvectors of the generalized eigensystem $Ax = tDx$. Then the fuzzy c-means clustering method is introduced into the d-dimensional space based on general Euclidean distance to cluster the data points. By maximizing the generalized modular function $\widetilde{Q}(U_d)$ for varying d, we obtain the appropriate number of clusters. The final soft assignment matrix determines the final clusters' membership with a designated threshold λ.

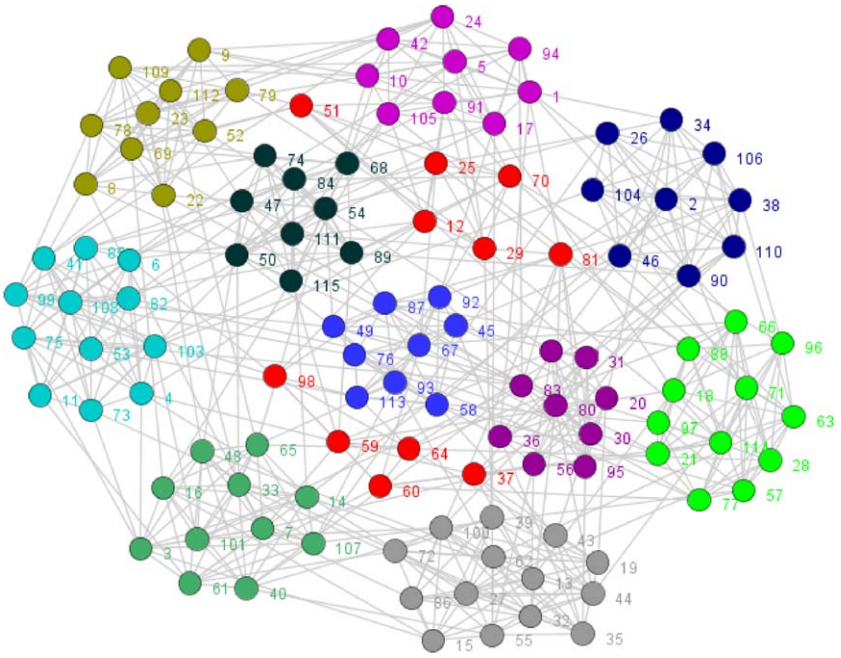


Fig. 5. Fuzzy communities of American college football team network ($k = 10$ and $\widetilde{Q} = 0.4673$) with given $\lambda = 0.10$ (best viewed in color).

Although spectral mapping has been comprehensively used before to detect communities in complex networks (even in clustering the real-valued points), we believe that our method represents a step forward in this field. A fuzzy method is introduced naturally with the generalized modular function and fuzzy c-means clustering technique. As our tests have suggested, it is very natural that some nodes should belong to more than one community. These nodes may play a special role in a complex network system. For example, in a biological network such as protein interaction network, one node (protein or gene) belonging to two functional modules may act as a bridge between them which transfers biological information or acts as multiple functional units [6].

One thing should be noted is that when this method is applied to large complex networks, computational complexity is a key problem. Fortunately, some fast techniques for solving eigensystem have been developed [20] and several methods of FCM acceleration can also be found in the literature [21]. For instance, if we adopt the implicitly restarted Lanczos method (IRLM) [20] to compute the $K-1$ eigenvectors and the efficient implementation of the FCM algorithm in Ref. [21], we can have the worse-case complexity of $O(mKh + nK^2h + K^3h)$ and $O(nK^2)$, respectively, where m is the number of edges in the network and h is the number of iteration required until convergence. For large sparse networks where $m \sim n$, and $K \ll n$, the algorithms will scale roughly linearly as a function of the number of nodes n. Nonetheless, the eigenvector computation is still the most computationally expensive step of the method.

We expect that this new method will be employed with promising results in the detection of communities in complex networks.

Acknowledgments

This work is partly supported by Important Research Direction Project of CAS "Some Important Problem in Bioinformatics", National Natural Science Foundation of China under Grant No.10471141. The authors thank Professor M.E.J. Newman for providing the data of karate club network and the college football team network.

References

[1] M.E.J. Newman, Detecting community structure in networks, Eur. Phys. J. B 38 (2004) 321–330.
[2] L. Danon, J. Duch, A. Diaz-Guilera, A. Arenas, Comparing community structure identification, J. Stat. Mech. P09008 (2005).
[3] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.
[4] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[5] S. White, P. Smyth, A spectral clustering approach to finding communities in graphs, SIAM International Conference on Data Mining, 2005.
[6] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.
[7] J. Reichardt, S. Bornholdt, Detecting fuzzy community structures in complex networks with a Potts model, Phys. Rev. Lett. 93 (2004) 218701.
[8] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[9] L. Donetti, M.A. Muñoz, Detecting network communities: a new systematic and efficient algorithm, J. Stat. Mech. P10012 (2004).
[10] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybernet. 3 (1973) 32–57.
[11] D. Verma, M. Meilă, A comparison of spectral clustering algorithms. Technical Report, 2003, UW CSE Technical Report 03-05-01.
[12] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, Adv. Neural Inf. Process. Systems 14 (2002) 849–856.
[13] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[14] Fuzzy Clustering Toolbox-⟨http://www.fmt.vein.hu/softcomp/fclusttoolbox/⟩.
[15] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (1977) 452–473.
[16] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (2004) 066133.
[17] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101 (9) (2004) 2658–2663.
[18] F. Wu, B.A. Huberman, Finding communities in linear time: a physics approach, Eur. Phys. J. B 38 (2004) 331–338.
[19] S. Fortunato, V. Latora, M. Marchiori, A method to find community structures based on information centrality, Phys. Rev. E 70 (2004) 056104.
[20] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. Vorst (Eds.), Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, SIAM, Philadelphia, PA, 2000.
[21] J.F. Kelen, T. Hutcheson, Reducing the time complexity of the fuzzy *c*-means algorithm, IEEE Trans. Fuzzy Systems 10 (2) (2002) 263–267.