Uncovering fuzzy community structure in complex networks

Shihua Zhang,[1,3,*] Rui-Sheng Wang,[2] and Xiang-Sun Zhang[1]
[1]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China*
[2]*School of Information, Renmin University of China, Beijing 100872, China*
[3]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China*
(Received 22 November 2006; revised manuscript received 9 June 2007; published 5 October 2007)

There has been an increasing interest in properties of complex networks, such as small-world property, power-law degree distribution, and network transitivity which seem to be common to many real world networks. In this study, a useful community detection method based on non-negative matrix factorization (NMF) technique is presented. Based on a popular modular function, a proper feature matrix from diffusion kernel and NMF algorithm, the presented method can detect an appropriate number of fuzzy communities in which a node may belong to more than one community. The distinguished characteristic of the method is its capability of quantifying how much a node belongs to a community. The quantification provides an absolute membership degree for each node to each community which can be employed to uncover fuzzy community structure. The computational results of the method on artificial and real networks confirm its ability.

DOI: 10.1103/PhysRevE.76.046103 PACS number(s): 89.75.Hc, 87.23.Ge

I. INTRODUCTION

Modularity or community structure is a natural characteristic in many real networks such as social networks [1,2], technological networks [3], and biological networks [4–7]. The detection of community structure in complex networks can enhance the insight into the intrinsical structure of networks and then has become a key problem in the study of networked systems.

Although the community structure, as a densely connected subgraph which sparsely connects with other parts of a network, is easily understood, giving out a deterministic definition is a nontrivial problem for the complexity of networks. A huge number of methods intended to detect the community structure in complex networks have been recently reviewed in [8] and evaluated in [9].

Recently, a concept of modularity Q introduced by Newman and Girvan [10] has been broadly used as a valid measure for community structure. In detail, given an undirected graph or network $G(V,E)$ consisting of the node set V and the edge set E, its adjacency matrix is denoted as $A=[a_{ij}]_{n\times n}$, where $a_{ij}=1$, if nodes i and j are connected and otherwise $a_{ij}=0$. Let n be the size of the node set. The modularity function Q is defined as

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

where P_k is a partition of the nodes into k groups and $L(V',V'')=\Sigma_{i\in V',j\in V''}a_{ij}$. The modularity function provides a way to determine if a partition is valid to decipher the community structure in a network. Maximization of the modularity function Q over all the possible partitions of a network is now a highly effective method [8–10]. Based on the modularity function, many methods have been developed, among which a recent breakthrough is made by Newman [11] in which a fast and accurate spectral algorithm has been developed.

An important case in community detection is that some nodes may not belong to a single community and then placing them into more than one group is more reasonable. Such nodes may mean a "fuzzy" categorization and take a special role such as signal transduction in biological networks. But the overlap of community structure cannot be detected by most existing partitioning algorithms and hierarchical clustering methods. Only a few community-detection methods can achieve this point [12–14]. Another phenomenon is that some nodes located on the border between two communities are hard to be classified into any community. Such nodes are considered as unstable nodes in Ref. [15] where the authors design an algorithm to identify the unstable nodes lying between two communities. Figure 1 shows a typical example where node 11 should be classified into two communities intuitively and node 6 lies exactly between two clear communities.

Here, we introduce a new community detection algorithm which can uncover meaningful fuzzy community structure in complex networks. The novel method can quantify the degree that each node belongs to each community. Based on the difference of membership degrees, we can uncover fuzzy communities in which a node may belong to more than one community. The algorithm does not need any prior knowledge about the number of communities and can give an appropriate number by maximizing the modular function. Applying the presented method to several artificial and real networks shows its capability.

II. NON-NEGATIVE MATRIX FACTORIZATION

Our research is motivated by the NMF technique, a machine-learning algorithm based on decomposition by parts that can uncover localized features in feature space [16,17]. The technique decomposes the feature matrix into two matrices with non-negativity constraints. And it was adapted to

*zsh@amss.ac.cn

 ©2007 The American Physical Society

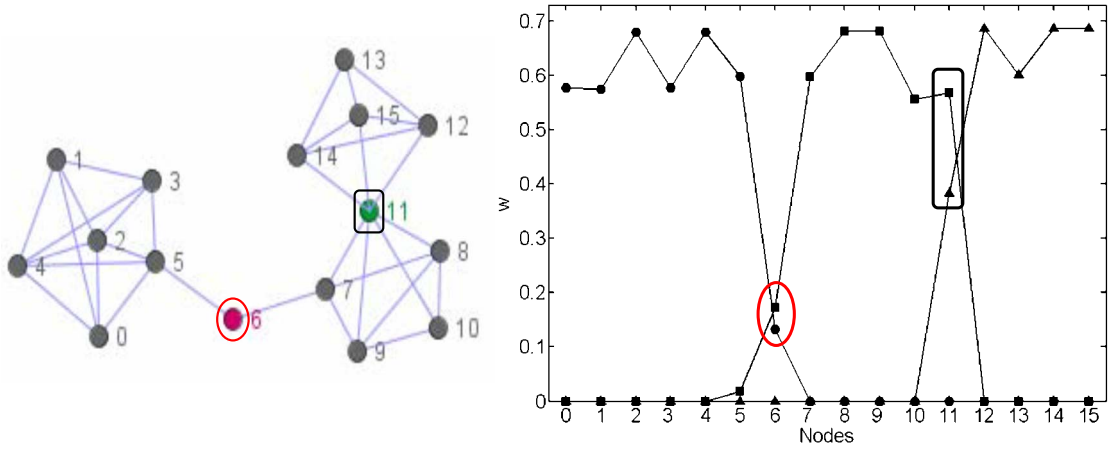

FIG. 1. (Color online) A toy network with one unstable node (6) and one overlapping node (11) and its corresponding W entries. The three curves (circle, square, and triangle) represent the corresponding values of every node in every column of W.

work as a clustering algorithm and dimensionality reduction technique in many fields [18,19].

Feature matrix. How should we determine a feature matrix to store the topological information of a network? Obviously, many approaches can be used. Here we employ the diffusion kernel [20] which has been comprehensively used in various fields [21]. Given an undirected, unweighted graph (network) $G=(V,E)$. The (opposite) Laplacian of this network is the following matrix:

$$L_{ij} = \begin{cases} 1, & \text{for } i \sim j \\ -d_i, & \text{for } i = j \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $i \sim j$ means that the ith and jth nodes are connected by an edge, and d_i is the degree of node i. The exponential of matrix L is defined as

$$K \equiv \exp(\beta L) = \lim_{n\to\infty}\left(1 + \frac{\beta L}{n}\right)^n, \tag{3}$$

where β is a positive constant to control the degree of diffusion. The resulting matrix K is symmetric and positive definite. It is naturally a valid kernel, which can capture the long-range relationship between nodes induced by the local structure of the network. As to the efficient computation of the exponential, many algorithms have been developed [22]. For example, the Padé approximation with scaling and squaring has been used to compute in MATLAB soft [23]. A similarity matrix B can be obtained by normalizing the kernel matrix K in such a way:

$$B_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}. \tag{4}$$

General approach. Feature data from a network is represented as a single matrix V of size $n \times m$ (in our study, it represents the symmetric matrix B, and so $n=m$). Generally, the column and row represent two different attributes, but here column and row both correspond to the similarities from one node to all nodes because of the symmetry of V. The major analytical method applied here, NMF, is an approximate factorization of the matrix V into a pair of matrices W and H,

$$V \simeq W \cdot H. \tag{5}$$

Note that this is only an approximate factorization, not an exact one [16,17]. That none of the matrices in this equation is permitted to have negative entries [17] is the unique feature of the NMF algorithm. The factorization is carried out with a particular rank k so that W is of dimension $n \times k$ and H is $k \times m$. Moreover, the factorization could be viewed as a representation of the data in a new space of lower dimensionality (k). Generally, there is a dual interpretation of decomposition. More interestingly, since the feature matrix V is symmetric, so $V=V^T \simeq H^T W$. W and H^T can be considered equivalent in a scale view. This has also been shown experimentally, so here we always employ W to determine the final clustering partition.

Implementation of NMF. The NMF algorithm is coded using the MATLAB version 6.5 [23]. It is the key of the algorithm to iteratively update matrices W and H to improve the approximation to V while maintaining non-negative matrix entries throughout [17]. For a given value of the NMF dimensionality k, the algorithm starts with random matrices W and H. The initial matrices for W and H with random entries are chosen from a normal distribution with mean 0, variance 1, and standard deviation 1. If an entry of the matrix is negative we take its absolute value to replace it. The two matrices are iteratively updated using the following rules:

$$H_{au} \leftarrow H_{au}\frac{(W^T V)_{au}}{(W^T W H)_{au}}, \tag{6}$$

$$W_{ia} \leftarrow W_{ia}\frac{(V H^T)_{ia}}{(W H H^T)_{ia}} \tag{7}$$

which minimize the root-mean-square (rms) error ($E = \|V - WH\|_2$) between the actual data V and the reduced-dimension reconstruction of the data WH. Because the up-

date rules are multiplicative, initial non-negative matrices remain non-negative for all the iterations. Furthermore, E is nonincreasing under the above update rules. Iteration continues until the absolute change of rms error is $<10^{-5}$ in an iteration or the total iterations attain a certain number, for example 200 steps. The update rules are a kind of gradient descent, and thus can only converge to a local minimum. For a k, given initial matrices W and H, we can find a good approximate factorization by running the iteratively updating procedure and stop criterion.

The factorization can be considered that each data vector v (the row of V) is approximated by a linear combination of the rows of H weighted by the components of w (the row of W): $v=wH$. Therefore H can be regarded as a basis that is optimized for the linear approximation of the feature data in V [16,17]. We can see that relatively few basis vectors are used to represent many data vectors and the entries of w represent the weight of every basis vector to produce the data vector v. Given a factorization $V \simeq WH$, we can use matrix W to group the n objects into k clusters. The entries in W can be viewed as the memberships of every node to each community. These memberships are absolute (i.e., not relative) and denote degrees of belonging or typicality. Unlike the fuzzy c-mean algorithm [14], the membership value of a node in one community does not depend on its membership values in other communities. Therefore a node can have high membership degrees in several communities and can also have very low membership degrees in all communities. So NMF can achieve our ideas about fuzzy or overlapping community structure.

Each object i will be placed into a cluster j^* if the w_{ij^*} is the largest entry in row i, i.e., $j^*=\arg\max_j w_{ij}$, so that the NMF algorithm can determine a partitioning clustering. We observed that one node may have strong association with more than one group which can be reflected by the association matrix W. So if the second largest association value is still large relative to the largest one, it means that the corresponding node is an unstable node. We introduce the following rule to depict the stability of nodes (called stable index, S):

$$S_i = \frac{w_{ij^*}}{w_{ij^{**}}}, \tag{8}$$

where $j^{**}=\arg\max_{j,j\neq j^*} w_{ij}$. We can find that the smaller the S_i, the more unstable is the node i. Figure 1 shows the corresponding w_{ij} values for every node i $(i=0,\ldots,15)$ and $j=1,2,3$. It is easy to find that node 6 and 11 have two prominent w values, respectively, and naturally very smaller S values. So, by our method we can uncover that these two nodes belong to two communities, respectively.

An important problem is how we can determine the value of k. Taking into account that the NMF method can produce hard clustering results, we employ an objective function such as Q modularity function or local modularity LQ measure [24] to determine an appropriate k value. Since NMF starts with random matrices, different implementations may return different results. We overcome this by repeating the algorithm for certain times, for example, 50 times, and select one solution using Q for a given k.

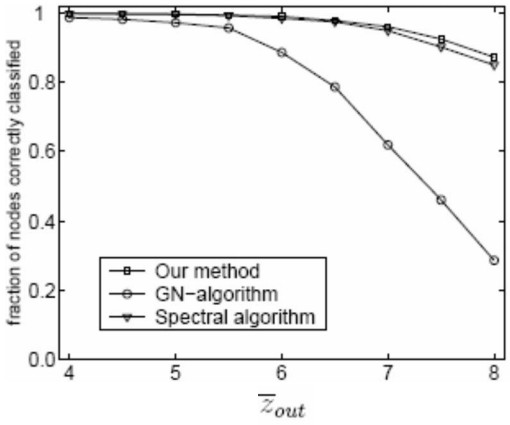

FIG. 2. Test of the method on computer-generated networks with known community structure and comparison with the GN algorithm and the spectral algorithm. It is a plot of the fraction of nodes correctly classified in computer-generated networks with respect to $\overline{z}_{out}$. Each point is an average over 100 realizations of the networks.

III. EXPERIMENT

We test the performance of the method proposed here by applying it to a class of artificial networks and to three real-world networks. Well-tried results and comparison with the known "GN algorithm" [25] and the new fast spectral algorithm [11] show the usefulness of the proposed method. And if there is no special mention, we choose $\beta=0.1$ in the feature matrices in our study.

A. Computer-generated networks

The NMF method for basic hard clustering is applied to a large set of artificial modular networks to compare with GN algorithm [25] and the spectral algorithm developed by Newman [11]. The experiment designed by Girvan and Newman [25] has been broadly used to test community-detection algorithms [25–29] in recent years. In this test, each network has 128 nodes, which are divided into four communities of size 32 each. Edges are placed randomly with two fixed expectation values so as to keep the average degree of a node to be 16 and the average $\overline{z}_{out}$ of each node's edges connecting to nodes of other modules. Obviously, as $\overline{z}_{out}$ increases, the classification of nodes becomes more and more difficult for any method.

The proposed method can effectively uncover the known four communities, i.e., $k=4$. If a node in a given community is classified into a detected group that contains the most nodes of this community, we consider it as a correctly classified node. Figure 2 shows the fraction of nodes that are classified into their correct communities with respect to $\overline{z}_{out}$ by our method, the GN algorithm, and the spectral algorithm. Our method has extremely better performance than the GN algorithm and comparative results with the spectral algo-

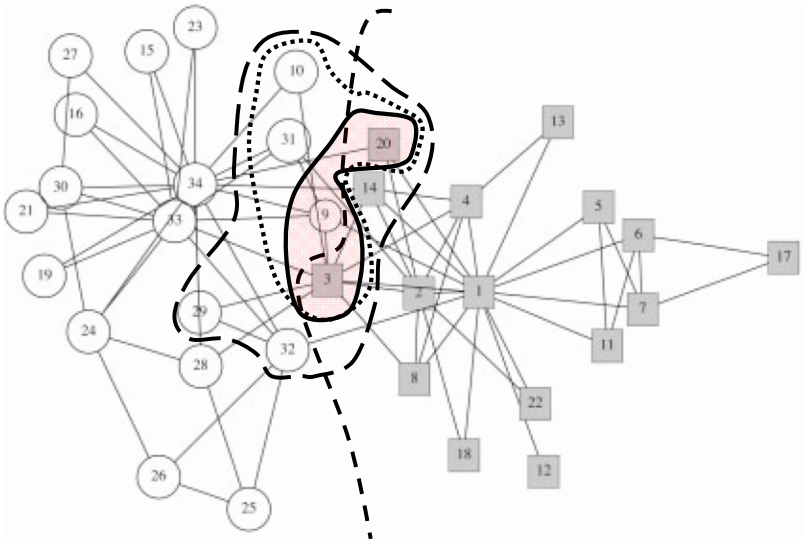

FIG. 3. (Color online) The fuzzy community structure of the karate club network detected by the proposed method.

rithm. For instance, for 100 random networks with $\bar{z}_{out}=7$, on an average 96.2% nodes are classified correctly by our method, while only about 61.9% nodes by the GN algorithm and about 95.0% nodes by the spectral algorithm.

B. The karate club network

The famous karate club network analyzed by Zachary [30] is widely used as a test example for methods of detecting communities in complex networks [8,25,26,31,32]. 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 concerned is 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. Figures 3 and 4 show the network and its corresponding results. Our NMF method employed as a hard-clustering algorithm divides the network into two groups of roughly equal size and produces a completely consistent split with the actual division of the original club. This indicates that the application of our method to the empirically observed network can uncover its real situation, and further we can detect some nodes belonging to more than one community which constitute the fuzzy boundaries of two communities. The three most unstable nodes including nodes 9, 3, 20 are depicted in the innermost bold loop region. These three nodes are exactly in-between nodes, between the two smaller clubs. This means that such members have good friendship with the two clubs at the same time. Also we can uncover more such nodes with a different degree of unstability according to the sorted S values. The five nodes with the five lowest S values are shown in the dotted loop regions. Figure 4 shows the change trend of stable index from which we can see that there is a distinct jump point which determines the outermost loop region consisting of eight nodes.

C. The scientific collaboration network

The scientific collaboration network collected by Girvan and Newman [25] and examined in [25,26] is also tested here. This network is a weighted network which consists of 118 nodes (scientists). The peak for scientific collaboration network is at $k=7$, $Q=0.7172$ (see Fig. 5). Figure 6 shows the fuzzy community structure detected by the proposed

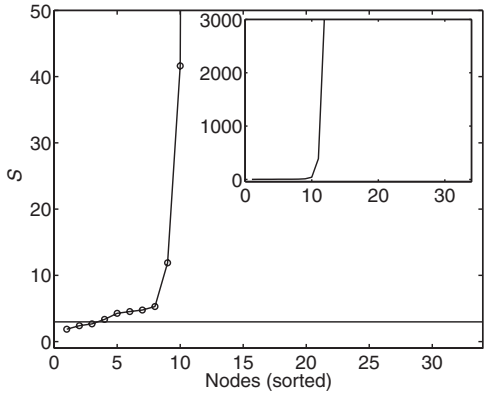

FIG. 4. The sorted S values for the karate club network. The inline small figure is plotted with larger S bound.

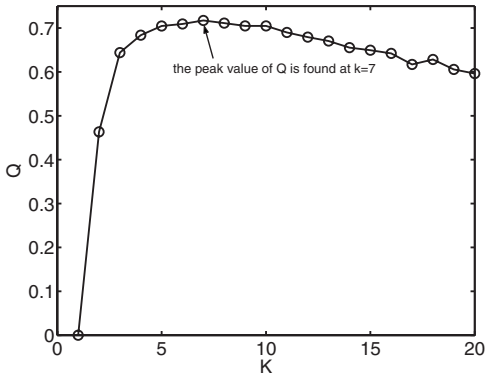

FIG. 5. Q values vs k of our method for scientific collaboration network.

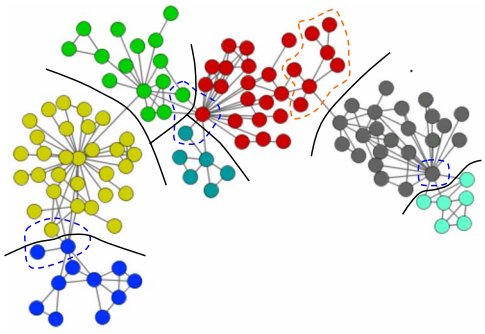

FIG. 6. (Color online) The fuzzy community structure of scientific collaboration network obtained by our method.

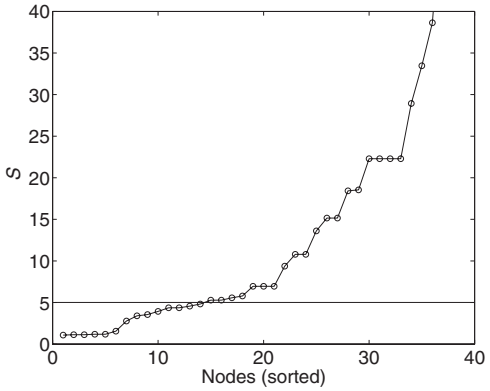

FIG. 7. The sorted S values for scientific collaboration network.

method which is visually very reasonable. Furthermore, four regions (including 14 overlapping nodes enclosed by four circles in Fig. 6) are detected according to their S values. These nodes generally locate on the borders of two or more communities and represent authors with multiple research interests or cross-discipline background. Maybe such points play a role in bridging two or more communities in a complex network of other types. The ability to find such nodes is a distinguished characteristic of our method.

Figure 7 shows the sorted stable index for scientific collaboration network. In view of the complexity of networks, giving out a criterion for choosing the threshold of S is a hard problem. In other words, the results of our method depend on the choice of S. When applied to real networks, the choice of S should rely on several trials and experiential knowledge. For example, we choose $S=2.4$ in the scientific collaboration network.

D. A large-scale protein interaction network

Large-scale yeast nonredundant (no self-interaction and repeated interaction) protein interaction data are obtained from [33] to construct a yeast protein interaction network which contains 2708 proteins (nodes) and 7123 interactions (edges). We apply the present method to this large-scale network to show its performance in uncovering communities (functional modules). The biological significance of these modules can be evaluated based on known function annotation and protein complexes in MIPS (Munich Information Center for Protein Sequences) database [34]. We can determine k (the number of communities in the network) based on the Q or LQ in a bisection manner. Here, we focus on checking the ability of the present method to detect fuzzy modules in large-scale networks, so we assume k is known. When $k=200$, we obtained 197 modules among which 196 ones are of sizes from 3 to 55 as well as a big one with 140 proteins (three small ones with two proteins are deleted). The nodes with 15% lowest S values are selected as overlapping nodes which constitute the fuzzy boundaries of overlapping modules. Figure 8 illustrates three fuzzy modules. The overlapping proteins in them may take special roles in signal transduction or communication among different functional modules.

Like most other community-detection algorithms, the main computational time of our method lies in searching a proper k. Besides this, the core computation is spent on the update rules where the matrix computation is relatively intense. The NMF method has the worse-case time complexity of $O(hkn^2)$ for a given k, where n is the number of nodes in the network and h is the number of iterations required until convergence. The computational time scales roughly quadratically as a function of the number of nodes n. Experimentally, computing an approximate factorization of feature matrix is fast and can be used to deal with networks with several thousands of nodes in several minutes. For the choice of k, the method can be easily performed in a parallel manner. Hence it is estimated that the NMF method can be ap-

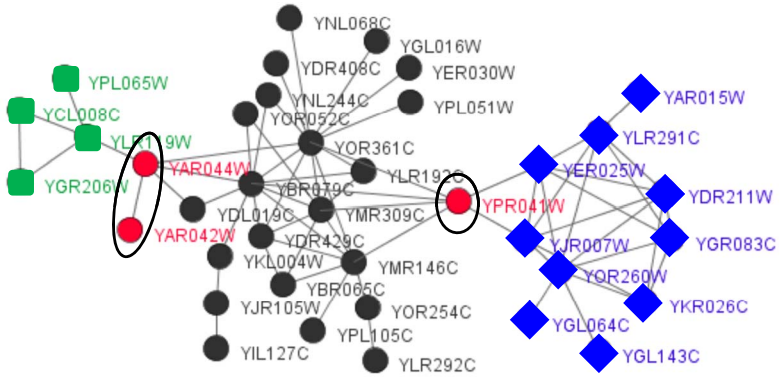

FIG. 8. (Color online) Three fuzzy modules (module-specific nodes are denoted by different shapes) have been shown with three overlapping proteins YAR044W, YAR042W and YPR041W (labeled in circles).

plied to real large networks with about tens of thousands of nodes.

IV. CONCLUSION AND DISCUSSION

In this paper, we present a method based on NMF technique to uncover fuzzy community structure in complex networks. 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 a 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 a multiple functional unit [13]. While the stable index proposed here is simple, it can afford abundant information about the organization of networks.

Although many community-detection algorithms have been developed before in the field of complex networks, only a few of them can detect "fuzzy" or "overlapping" community structure [12–14]. The clique percolation method (CPM) [13] detects communities based on a class of basic elements—cliques. It is too restrictive and only a few communities can be detected with many nodes excluded, especially in sparse networks [35]. The Potts model for fuzzy community detection [12] is a random search procedure and returns different assignments of nodes upon different initial assignments. They repeat the algorithm many times and combine the inconsistence of these assignments to form fuzzy communities. The fuzzy clustering method (FCM) used in our recent work [14] can only give a relative membership, while the new NMF method quantifies how much a node belongs to a community and employs an absolute membership in an elaborate mathematical manner which is more reasonable since it can reflect the absolute possibility that a node belongs to a specific community. This point is somewhat related to the studies in Refs. [36,37], in which they defined two indexes to describe different roles of nodes according to their pattern of within- and between-module connections. Obviously, the quantification of the degree that a node belongs to a community can be employed to do a similar study. Therefore though there have been many algorithms for detecting community structure, clearly the method in this paper can be a helpful complement to the existing ones.

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

ACKNOWLEDGMENTS

This work was partly supported by Important Research Direction Project of CAS Some Important Problems in Bioinformatics, the National Natural Science Foundation of China under Grant No. 10631070, No. 10701080, and the Ministry of Science and Technology, China, under Grant No. 2006CB503905. The authors thank Professor M. E. J. Newman for providing the data of karate club network, SFI collaboration network, and the source code of the new spectral algorithm.

[1] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[2] A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[3] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, IEEE Computer **35**, 66 (2002).
[4] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A.-L. Barabasi, Nature (London) **407**, 651 (2000).
[5] A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001).
[6] H. Jeong, S. Mason, A.-L. Barabasi, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).
[7] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi, Science **297**, 1551 (2002).
[8] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).
[9] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas, J. Stat. Mech.: Theory Exp. 2005, P09008 (2005).
[10] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[11] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[12] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[13] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[14] S. Zhang, R. S. Wang, and X. S. Zhang, Physica A **374**, 483 (2007).
[15] D. Gfeller, J. C. Chappelier, and P. De Los Rios, Phys. Rev. E **72**, 056135 (2005).
[16] D. D. Lee and H. S. Seung, Nature (London) **401**, 788 (1999).
[17] D. D. Lee and H. S. Seung, Adv. Neural Inf. Process. Syst. **13**, 556 (2001).
[18] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, Proc. Natl. Acad. Sci. U.S.A. **101**, 4164 (2004).
[19] W. Liu and N. Zheng, Pattern Recogn. Lett. **25**, 893 (2004).
[20] R. I. Kondor and J. Lafferty, in Proceedings of the Nineteenth International Conference on Machine Learning (Morgan Kaufmann, San Francisco, 2002), pp. 315–322.
[21] R. Kondor and J.-P. Vert, in *Kernel Methods in Computational Biology*, edited by B. Scholkopf, K. Tsuda, and J.-P. Vert (MIT Press, Cambridge, MA, 2004).
[22] C. B. Moler and C. F. Van Loan, SIAM Rev. **20**, 801 (1979).
[23] http://www.mathworks.com/
[24] S. Muff, F. Rao, and A. Caflisch, Phys. Rev. E **72**, 056107 (2005).
[25] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[26] F. Wu and B. A. Huberman, Eur. Phys. J. B **38**, 331 (2004).
[27] H. J. Zhou, Phys. Rev. E **67**, 061901 (2003).
[28] S. Zhang, X. M. Ning, and X. S. Zhang, Eur. Phys. J. B **57**, 67 (2007).
[29] S. Fortunato, V. Latora, and M. Marchiori, Phys. Rev. E **70**,

056104 (2004).
[30] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[31] L. Donetti and M. A. Muñoz, J. Stat. Mech.: Theory Exp. 2004, P10012 (2004).
[32] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[33] N. J. Krogan *et al.*, Nature (London) **440**, 637 (2006).
[34] H. W. Mewes *et al.*, Nucleic Acids Res. **30**, 31 (2002).
[35] S. Zhang, H. W. Liu, X. M. Ning, and X. S. Zhang, in Proceedings of IEEE International Conference on Data Mining-Workshops (ICDMW'06) (IEEE Computer Society, Washington D.C., 2006), pp. 130–135.
[36] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[37] R. Guimerà and L. A. N. Amaral, J. Stat. Mech.: Theory Exp. 2005, P02001 (2005).