Eur. Phys. J. B **57**, 67–74 (2007)
DOI: 10.1140/epjb/e2007-00146-y

THE EUROPEAN PHYSICAL JOURNAL B

Graph kernels, hierarchical clustering, and network community structure: experiments and comparative analysis

S. Zhang[1,2,a], X.-M. Ning[1,2], and X.-S. Zhang[1]

[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, P.R. China

Received 12 July 2006 / Received in final form 20 April 2007
Published online 25 May 2007 –

Abstract. There has been a quickly growing interest in properties of complex networks, such as the small world property, power-law degree distribution, network transitivity, and community structure, which seem to be common to many real world networks. In this study, we consider the community property which is also found in many real networks. Based on the diffusion kernels of networks, a hierarchical clustering approach is proposed to uncover the community structure of different extent of complex networks. We test the method on some networks with known community structures and find that it can detect significant community structure in these networks. Comparison with related methods shows the effectiveness of the method.

PACS. 89.75.Hc Networks and genealogical trees – 89.65.-s Social and economic systems – 05.10.-a Computational methods in statistical physics and nonlinear dynamics

1 Introduction

Since diverse systems in various fields take the form of networks, a number of recent studies have focused on several distinctive statistical properties of networks such as the small world property [1,2], the right-skewed degree distribution [3–5], the clustering or network transitivity property [6], the community structure [7,13] and so on. In this study, we proposed a simple approach to depict the community structure which is shared by many network systems.

Recently, modular organization of complex networks, as a basic feature for many real networks, has attracted special attention of researchers in diverse fields such as social networks [14,15], technological networks [16], and biological networks [17,18,21,22], etc. For example, many biological networks appear to be organized into community (modularity) structure that are densely connected within themselves but sparsely connected with the rest of the network. A large number of methods have been developed for this problem, such as the edge-betweenness algorithm [7], EO-based algorithm [8], and so on. More methods can be seen in a recent review article [9] and an evaluation paper [10]. The 'fuzzy' community structure also has been studied [11,12] recently. Many existing methods only take account of local information of each node, such as number of nearest neighbors shared with other nodes. Hierarchical clustering method based on similarity measurement between pairs of nodes according to the network structure as proposed in [17,24,23] is a direct and visualizing approach which detects community structure in network.

In addition, biological network tends to form a hierarchical structure, in which nodes are organized into small modules which are, in turn, organized into larger modules, and so on [17]. Hierarchical organization in complex networks is a key theoretical model which captures the statistical characteristics of a large amount of real networks such as metabolic networks [17,18], social networks [14,15]. Obviously, how to uncover the hierarchical structure in a network is a key, but nontrivial problem. Rarasz et al. (2002) remarked that conventional network clustering methods are unable to uncover the hierarchical structure in such networks. They have proposed a hierarchical clustering method to identify hierarchical modularity in metabolic network based on a so-called topological overlap matrix. In papers [19,20], the authors defined two distance measures based on network random walks respectively to partition the vertices into communities hierarchically.

As we know, until now, there is no a definite definition of the community structure. A subnet detected as a community by a constrained community-detection method may be only a part of a large community detected by another relaxed method. Giving out communities of

[a] Corresponding author: e-mail: zsh@amss.ac.cn

networks of different extent can enhance the insight into organization of complex networks.

In recent years, kernel method is becoming a popular tool in many fields including bioinformatics. Kernel function defines similarities between pairs of nodes and yields a symmetric, positive semidefinite matrix K known as the kernel matrix. Such similarities describe relationships that are implicit in the data and make them explicit. In order to decipher the topological link relationship of graph data, some graph kernels have been developed [25]. Recently, a novel graph kernel called diffusion kernel has been comprehensively used in many aspects such as data integration [26]. Naturally, it can be used as a similarity metric for hierarchical clustering and further detecting community structure in complex networks. A distinguished merit of this method is that it can control the degree of similarity based on the diffusion degree.

Here, we propose a simple hierarchical clustering method based on the concept of diffusion kernel of networks to uncover the community structure of complex networks. Applying this method to several artificial and real networks show the effectiveness of this method. Communities of different extent (relaxed or constrained) in a network enrich the understanding about the inherent structure of networks. Comparison shows some special advantages of the present method over the related methods.

2 Graph kernels and similarity index

A key factor of hierarchical clustering is the choice of dissimilarity metric (or similarity metric). For a network, it is a hard problem. A proper similarity measure S_{ij} for every pair of nodes i and j in a network should represent how closely connected the nodes are. Since two closely connected nodes would be in a close hierarchical level. Recently, Kondor et al. (2002) proposed a concept of diffusion kernel which implicitly captures the connecting relationship of nodes.

Given an undirected, unweighted graph (network) $G = (V, E)$. The (opposite) Laplacian of this network is the matrix:

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

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

$$K^\beta \equiv exp(\beta L) = \lim_{n\to\infty} \left(I + \frac{\beta L}{n}\right)^n \quad (2)$$

where β is a positive constant to control the degree of diffusion. And the limit always exists and is equivalent to the following expansion:

$$exp(\beta L) = I + \beta L + \frac{\beta^2}{2}L^2 + \frac{\beta^3}{3!}L^3 + \cdots \quad (3)$$

The resulting matrix K^β is symmetric and positive definite. It is therefore a valid kernel, which captures the long-range relationship between nodes induced by the local structure of the network. How should we compute the matrix exponential? As a matter of fact, many algorithms have been developed for this problem [27]. For example, the Padé approximation with scaling and squaring has been used to compute the matrix exponential in Matlab soft [28]. By normalizing the kernel matrix K^β, the similarity matrix S^β can be defined as:

$$S^\beta_{ij} = \frac{K^\beta_{ij}}{\sqrt{K^\beta_{ii}K^\beta_{jj}}}. \quad (4)$$

We note that the diffusion parameter β plays a key role in detecting community structure of different extent. Why does it runs? Thinking of diffusion kernel K in terms of actual physical process of diffusion can give some intuitional explanation. The parameter β of K is to control the extent of the diffusion, or to specify the length scale, which is similar to σ in the Gaussian kernel $e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}}$ [29]. The authors of reference [26] has suggested that diffusion kernel has closely relationship to random walk. A lazy random walk on an unweighted graph G is a stochastic process which generates sequences $z_0, z_1, z_2, ...$, where $z_l \in V$ in such a way that $p(z_{l+1} = j | z_l = i) = \beta_0$ (β_0 is a constant and $\beta_0 \leq 1/\max_i d_i$) if $i \sim j$ and zero otherwise, and remains in place with probability $1 - d_i\beta$. Obviously, the limit distribution $P = \lim_{N\to\infty} p(z_N|z_0)$ equals to $\lim_{N\to\infty}(I + \beta_0 L)^N$. Let $N = 1/\Delta t$ and $\beta_0 = \beta\Delta t$, we can easily obtain that the continuous time limit of lazy random walks are exactly the diffusion kernels. In a recent study, Zhou [19] has already given a dissimilarity measure based on network random walk in a very different manner. We should also note that the flexibility from the choice of β may mean a shortcoming at the same time. How should we determine an appropriate β for a large network? We suggest that two criteria used in [30] are adopted just as shown in part IV. B. *4. the yeast protein interaction network*. Experientially, the β is not sensitive to the two criteria. So we can only test several β to choose an appropriate β value. This will not increase the complexity of the algorithm.

3 Hierarchical clustering based on diffusion kernel (DK)

The normalized diffusion kernel can be used as a similarity index to decipher the community structure of a network. Generally, hierarchical clustering as a standard clustering technique can be used to cluster a network with a given similarity index. Starting from N clusters consisting of single node, the two closest ones are iteratively joined together. Three different criteria called 'single-linkage clustering' (SL), 'complete linkage clustering' (CL) and 'average-linkage clustering' (AL) can be employed to define cluster-to-cluster dissimilariy (see Fig. 1). Although these criteria have been broadly studied, none of them can be proved to be always more efficient than the

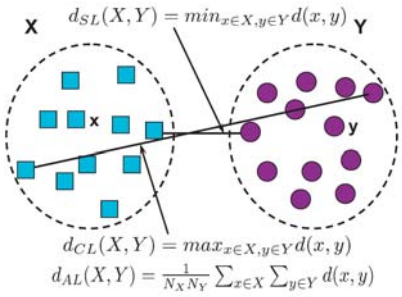

Fig. 1. The illustration of three cluster-to-cluster dissimilarity criteria. X and Y are two clusters and N_X, N_Y are the sizes of these two clusters. $d(x, y)$ is distance of $x \in X$ and $y \in Y$.

others. Taking into account the tendency to cluster the nodes together at a relatively low level and at a relatively high level for 'single-linkage clustering' and 'complete linkage clustering' respectively, we employ the compromising 'average-linkage clustering' in this study.

The output of hierarchical clustering can be depicted by a hierarchical tree or dendrogram. The whole procedure could be easily implemented with Fortran programming language. The hierarchical tree was displayed using TreeView [31] (`http://rana.lbl.gov/EisenSoftware.htm`). And if there no special mentions, the height of node of a hierarchical tree represents the similarity between two branches under the node in our paper.

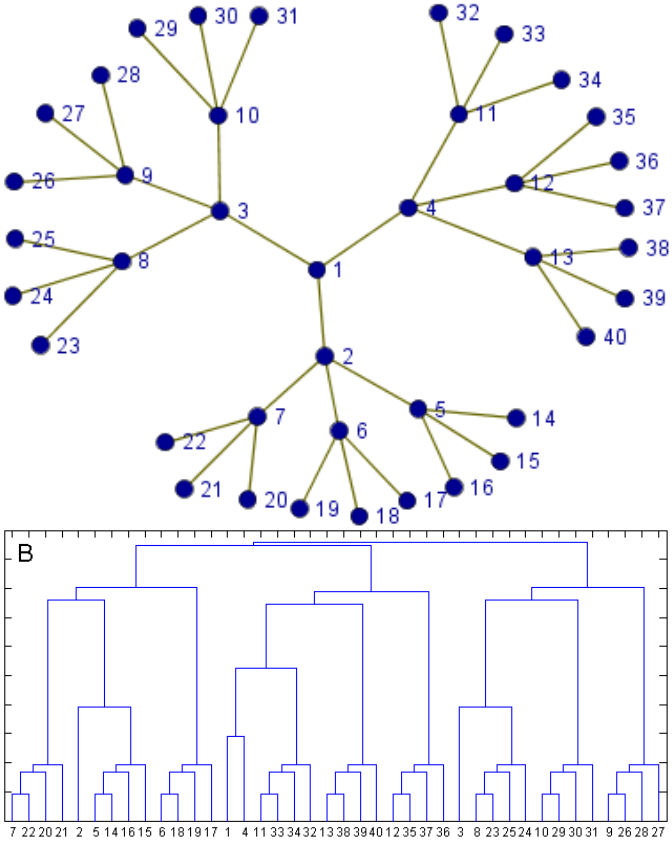

Fig. 2. The Cayley tree network and its hierarchical trees with $\beta = 0.3$.

4 Experiments and comparative analysis

We test the simple method by applying it to three kinds of artificial networks and to four real-world networks. Comparison with the known 'edge-betweenness' algorithm [7], the method proposed by [19] and other hierarchical clustering methods shows the effectiveness of our method.

4.1 Artificial networks

4.1.1 The Cayley-tree network

We first take a Cayley-tree network whose hierarchical structure is distinct as an example. Figure 2 illustrates the network and its tree-like plot by our method. We can see that the basic community structure and hierarchy structure are clear. The network can be divided into three parts at a high level. Furthermore, each of the three parts is further divided into three subparts.

4.1.2 The model hierarchy network

We apply the method to the model hierarchy network proposed by Ravasz et al. [17]. Such network can be constructed by a repeated replication and connection process from a four clique (Detailed information can be seen in [17]). Figure 3 shows such a network at level 2. As Ravasz et al. pointed, conventional network clustering methods are hard to uncover hierarchical community structure of such a network. The simple network method presented here makes a good performance. The global hierarchy organization of the nodes in the network has well reflection. Figure 4 demonstrates the community structure of the network in Figure 3 with $\beta = 3$.

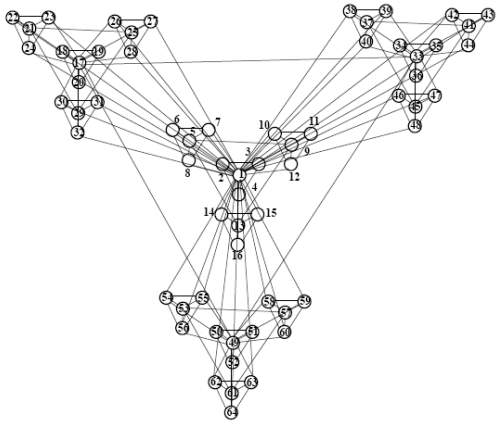

Fig. 3. A model hierarchy network obtained from [19].

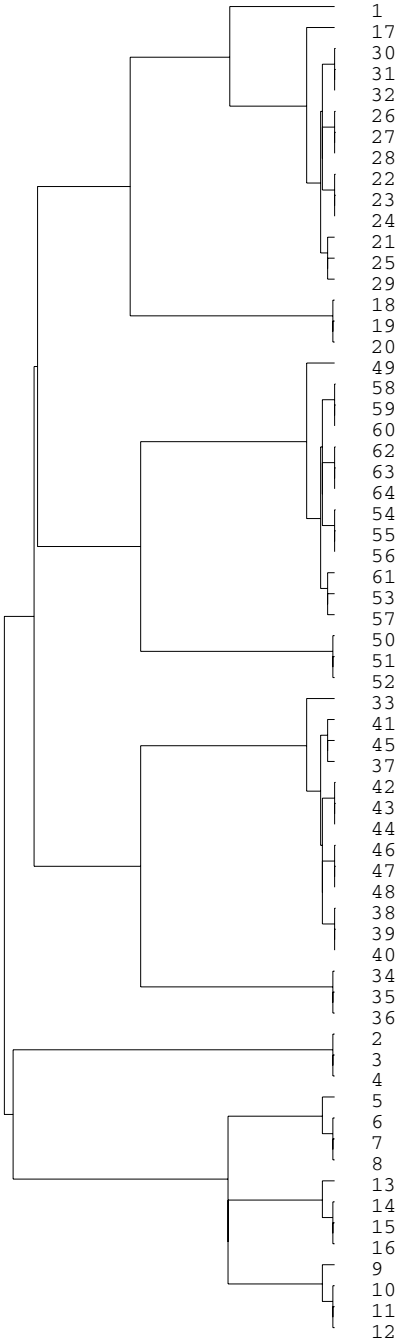

Fig. 4. The hierarchical trees of hierarchy network in Figure 3 with $\beta = 3$.

4.1.3 Computer-generated networks

The present method is applied to a large set of artificial modular networks to compare with edge-betweenness algorithm [7]. In this test, each network has 128 nodes, which are divided into 4 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. The experiment designed by

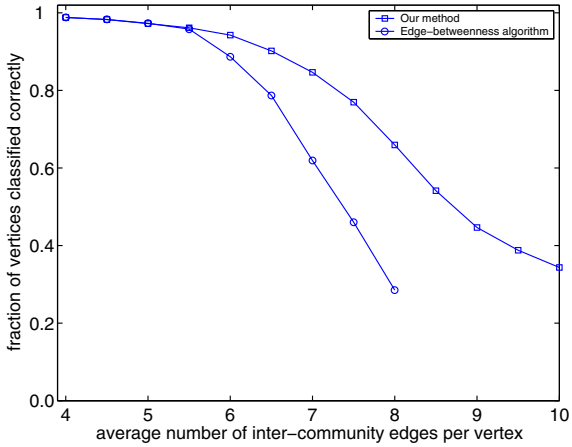

Fig. 5. The fraction of nodes correctly classified in computer-generated networks with respect to $\overline{z}_{out}$. Each point is an average over 100 realization of the networks.

Girvan and Newman [7] has been broadly used to test community-detection algorithms [7,32].

The proposed method can uncover the community structure well. Figure 5 shows the fraction of nodes that are classified into their correct communities with respect to $\overline{z}_{out}$ by our method with $\beta = 0.1$ and the edge-betweenness algorithm respectively. The present method has better performance than edge-betweenness algorithm. For instance, when worked on 100 random networks with $\overline{z}_{out} = 7$ by the present method, on an average 84.7% nodes are classified correctly, while only about 61.9% nodes by the edge-betweenness algorithm.

4.2 Real-world networks

4.2.1 The karate club network

The famous karate club network analyzed by Zachary [33] is widely used as a test example for methods of detecting communities in complex networks [7,9,13,32,34]. 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. Figure 6A shows the network, and Figures 6B and 6C show the hierarchical tree of communities produced by the edge-betweenness algorithm and our method with $\beta = 0.1$, respectively. Two methods both divide the network into two groups of roughly equal size at the top of the tree. Both methods produce almost consistent split with actual division of original club. And only one node, node 3 in Figure 6B and node 10 in Figure 6C respectively, is classified incorrectly. This indicates that the application of our direct hierarchical clustering method to the empirically observed network can uncover

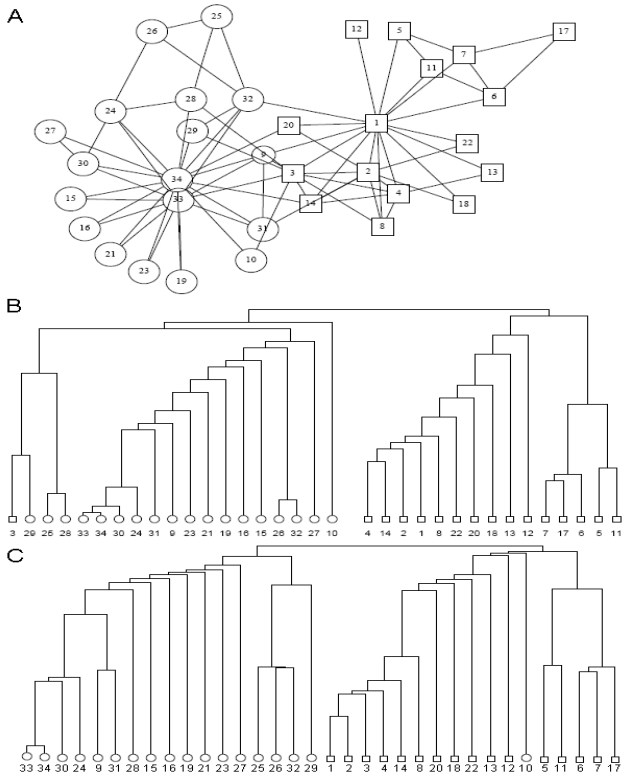

Fig. 6. (A) The karate club network from Zachary's study as described in the text (obtained from [9]). (B) (obtained from [7]) and (C) show the hierarchical trees by the edge-betweenness algorithm and our method with $\beta = 0.1$ respectively.

its roughly real hierarchy. But there exist more strongly connected subnets which are not detected well. Figure 7 shows two hierarchical trees, in which the left one is for our method with $\beta = 3$ and the right one is for method of Zhou [19] respectively. We can find several strongly connected subnets such as the one consisting of nodes 5, 6, 7, 11, 17. Specifically, the main three parts of the two plots in Figure 7 are quite consistent with each other except node 12.

4.2.2 The football team network

The second real network we have investigated is the college football network which represents the game schedule of the 2000 season of Divison 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 [7,19,32].

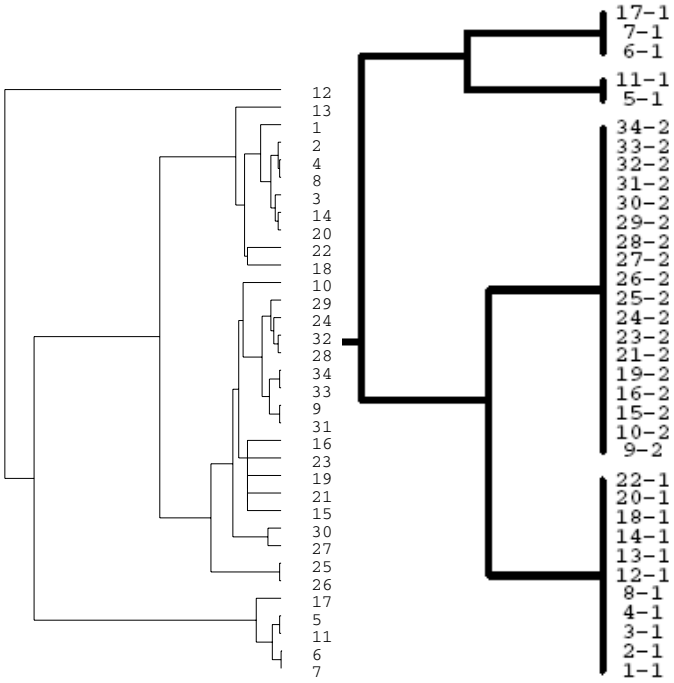

Fig. 7. The hierarchical structure of the karate club network obtained form the present method with $\beta = 3$ (the left one) and method developed by Zhou [19] (the right one which is obtained from [19]).

Figure 8 shows the community structure of the football team network calculated by using the present method (the left one) and obtained from [19] (the right one). The communities detected by both methods show well consistent with the nodes' group-identity. Because there are few edges between five members of the community labeled 12, these five nodes are distributed to other communities such as node 91 or deposited as relatively isolated node such as nodes 37 and 43. And the remainders of community 12 (nodes 81 and 83) joint together with community 8 at a high level. We should note that nodes including 37, 43, 81 and 83 may mean a unstable case in which these nodes only have weakly connecting relationship with any community. Nodes 59 and 111 are classified to inconsistent communities respectively for stronger link with current communities than the labeled ones. The hierarchical structure of our method seems to suggest a more precise organization than its original conferences.

4.2.3 The scientific collaboration network

The scientific collaboration network collected by Girvan and Newman [7] and examined in [7,32] is also tested here. This network is a weighted network which consists of 118 nodes (scientists). The proposed method gives out distinct community structure of different degree shown in Figure 9. The network can be divided into three giant communities, and small communities can further be detected just as method of Zhou [19] did.

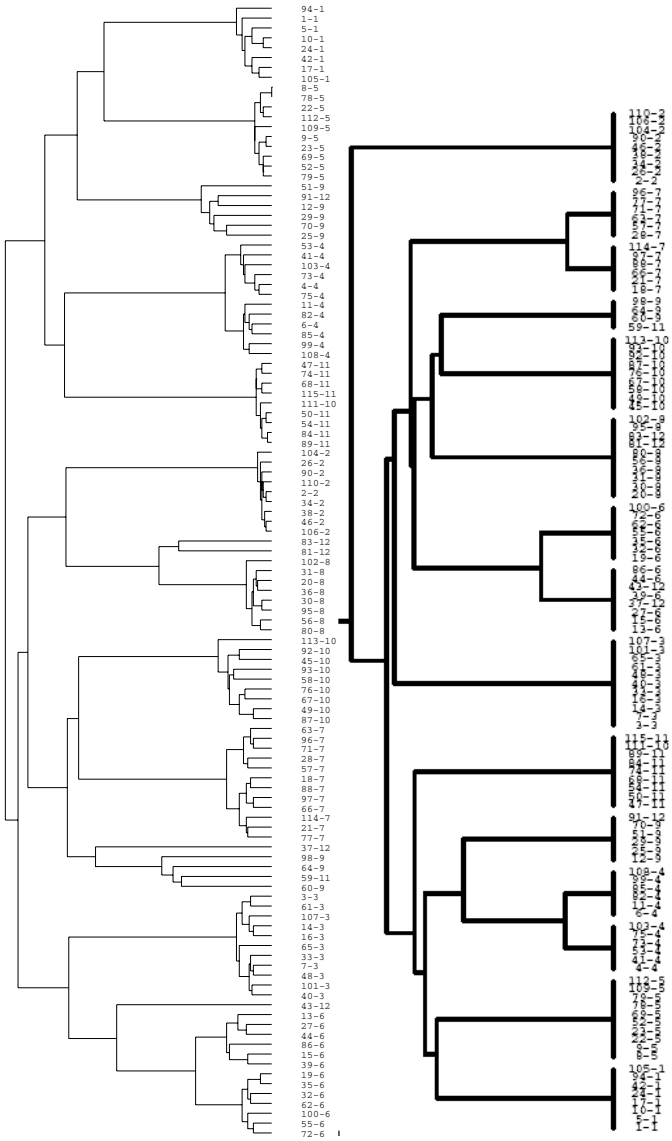

Fig. 8. The hierarchical structure of football team network obtained by our method with $\beta = 0.5$ (the left one) and the method prosed by Zhou [19] (the right one which is obtained from [19]).

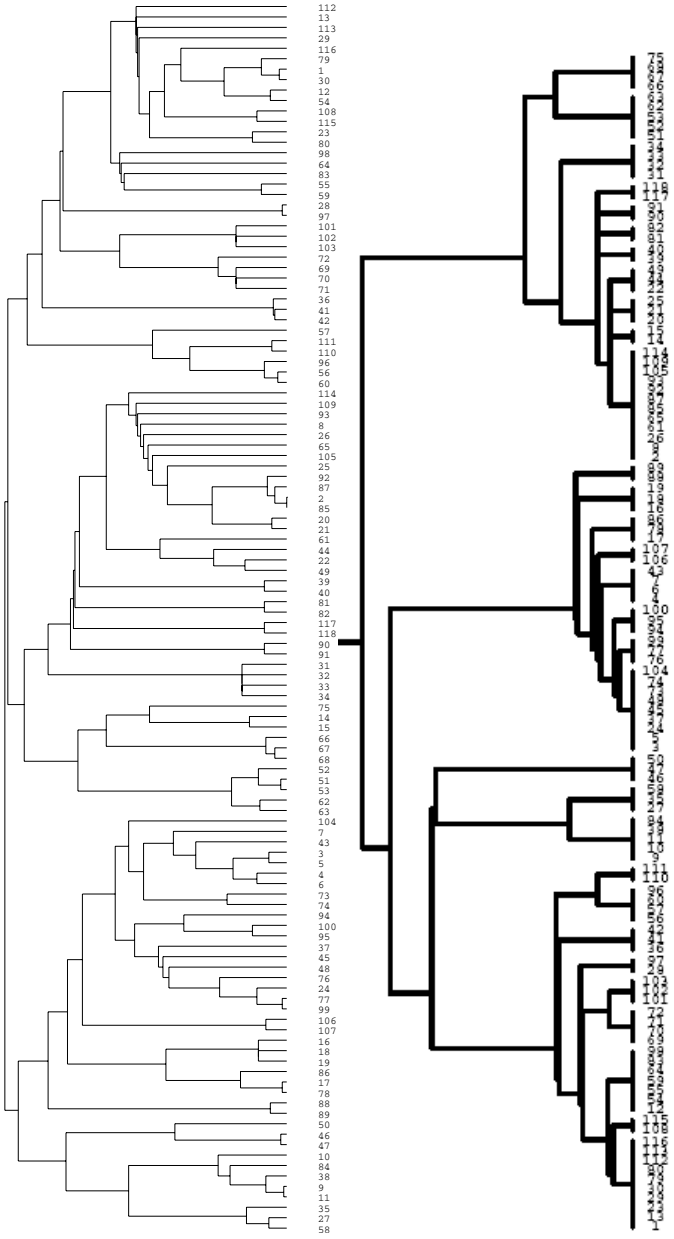

Fig. 9. The hierarchical trees of scientific collaboration network obtained by our method with $\beta = 3$ (the left one) and the method proposed by Zhou [19] (the right one which is obtained from [19]).

4.2.4 The yeast protein interaction network

A large-scale yeast nonredundant (no self-interaction and repeated interaction) protein interaction data is obtained from [30] to construct a yeast protein interaction network which contains 4537 proteins (nodes) and 13 344 interactions (edges). We apply the simple clustering method to this large sparse network. Many communities can be obtained. Its biological significance also can be evaluated based on known function annotation and protein complexes in MIPS [35].

In order to illustrate the advantages of the present method (called DK), we compared it with previously reported hierarchical methods based on different similarity measures including R [23], B [24], Ra [17]. Two criteria suggested in [30] are used to compare these methods. The first is to test how well the link information is retained in the dendrogram by computing the distribution of the shortest path in the tree between linked nodes. The second is to test how the nonzero entries in the adjacency matrix is ordered according to the hierarchical tree close to the diagonal of the matrix by computing the number of linked pairs in the selected area. Figures 10A and 10B show the shortest path distributions of linked nodes of the four trees and the number of linked pairs in selected area. We can see that the diffusion kernel-based method always shows better results than the others. These two

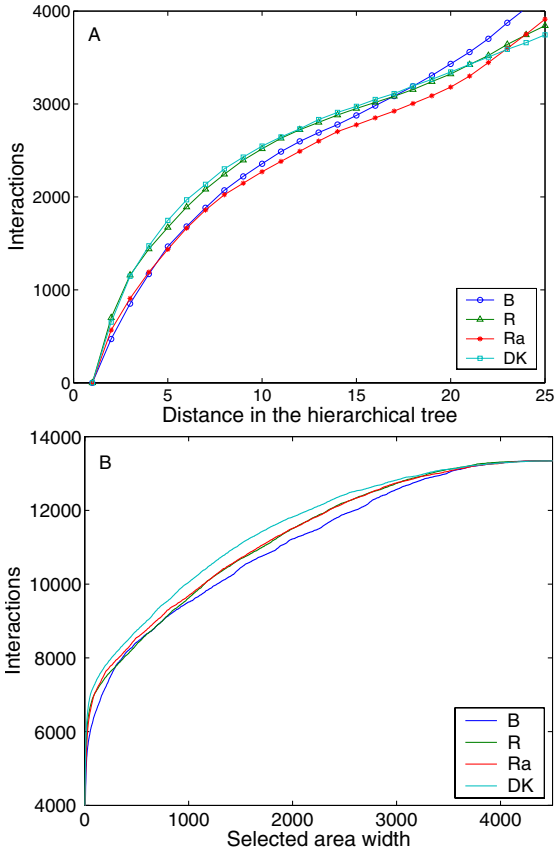

Fig. 10. (A) Distribution of interacting proteins according to the shortest path between them in the hierarchical tree with $\beta = 0.3$. (B) Distribution of protein interactions according to the selected area in the ordered adjacency matrix based on hierarchical tree with $\beta = 0.3$.

criteria can also be used as measures to choose appropriate β visually. In other words, we can compute these two distributions of diffusion kernels with an array of β values to choose a 'good' β value.

5 Conclusion and discussion

Graph kernel as an implicit similarity index provide an important relationship-depicting method which has been broadly applied to various fields such as bioinformatics. In this study, based upon diffusion kernels of networks we propose a hierarchical clustering method to uncover the community structure of complex networks. The hierarchical clustering method based on the kernel matrix is similar with traditional hierarchical clustering methods used in network clustering, but it is more flexible with the parameter β which controls the degree of diffusion and further controls the extent of communities. We apply this method to three artificial networks and four real networks in social and biological fields. The experiments show very satisfactory results. Comparison with other related methods shows the effectiveness of the present method. The different extent of communities provides enrich insights into network systems with diffusion parameter β just as the analysis on karate club network suggested. For very large networks, an appropriate β can be given out according to the criteria used in the analysis of protein interaction network.

Since diffusion kernels of unweighted networks can be generalized to weighted networks, we believe that our analysis can also be extended to weighted complex networks just as the test on the scientific collaboration network shown above. It is expected that the idea and approach presented here will be proved useful in the analysis of various types of complex networks.

This work is partly supported by Important Research Direction Project of CAS "Some Important Problems in Bioinformatics", National Natural Science Foundation of China under Grant No. 10631070. The authors thank Professor M.E.J. Newman for providing the data of karate club network, the college football team network and SFI collaboration network.

References

1. I. de S. Pool, M. Kochen, Social Networks **1**, 1 (1978)
2. S. Milgram, Psychol. Today **2**, 60 (1967)
3. A.-L. Barabsi, R. Albert, Science **286**, 509 (1999)
4. P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000)
5. S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000)
6. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
7. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002)
8. J. Duch, A. Arenas, Phys. Rev. E, **72**, 027104 (2005)
9. M.E.J. Newman, Eur. Phys. J. B **38**, 321 (2004)
10. L. Danon, J. Duch, A. Diaz-Guilera, A. Arenas, J. Stat. Mech. P09008 (2005)
11. J. Reichardt, S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004)
12. S. Zhang, R.S. Wang, X.S. Zhang, Physica A **374**, 483 (2007)
13. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Proc. Natl Acad. Sci., USA **101**, 2658 (2004)
14. M.E.J. Newman, Phys. Rev. E **64**, 016131 (2001)
15. A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek, Physica A **311**, 590 (2002)
16. G.W. Flake, S. Lawrence, C. Lee Giles, F.M. Coetzee, IEEE Computer **35**, 66 (2002)
17. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabasi, Science **297**, 1551 (2002)
18. H. Jeong, B. Tombor, R. Albert, Z. Oltvai, A.-L. Barabasi, Nature **407**, 651 (2000)
19. H.J. Zhou, Phys. Rev. E **67**, 061901 (2003)
20. H.J. Zhou, R. Lipowsky, *Proceeding of ICCS 2004*, edited by M. Bubak, G.D. Albada, P.M.A. Sloot, J. Dongarra, LNCS **3038**, 1062 (2004)
21. A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001)
22. H. Jeong, S. Mason, A.-L. Barabasi, Z.N. Oltvai, Nature **411**, 41 (2001)

23. A.W. Rives, T. Galitski, Proc. Natl Acad. Sci. USA **100**, 1128 (2003)
24. C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, B. Jacq, Genome Biol. **5**, R6 (2003)
25. F. Fouss, L. Yen, A. Pirotte, M. Saerens, *An experimental investigation of five graph kernels on a collaborative recommendation task*. (submitted for publication 2006)
26. R.I. Kondor, J. Lafferty, Internaltional Conference on Machine Learning (ICML2002) (2002)
27. C.B. Moler, C.F. Van Loan, SIAM Rev. **20**, 801 (1979)
28. `http://www.mathworks.com/`
29. R. Kondor, J.-P. Vert, in *Kernel Methods in Computational Biology*, edited by B. Scholkopf, K. Tsuda, J.-P. Vert (The MIT Press, 2004)
30. H. Lu, X. Zhu, H. Liu, G. Skogerbo, J. Zhang, Y. Zhang, L. Cai, Y. Zhao, S. Sun, J. Xu, D. Bu, R. Chen, Nucleic Acids Res. **32**, 4804 (2004)
31. M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Proc. Natl. Acad. Sci. USA **95**, 14863 (1998)
32. F. Wu, B.A. Huberman, Eur. Phys. J. B **38**, 331 (2004)
33. W.W. Zachary. J. Anthropol. Res. **33**, 452 (1977)
34. L. Donetti, Miguel, A. Muñoz, J. Stat. Mech, P10012 (2004)
35. H.W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, B. Weil, Nucleic Acids Res. **30**, 34 (2002)