Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Coarse-grained diffusion distance for community structure detection in complex networks

Jian Liu[1,3] **and Tingzhan Liu**[2]

[1] Key Laboratory of Mathematics and Applied Mathematics and School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China
[2] School of Sciences, Communication University of China, Beijing 100024, People's Republic of China
E-mail: dugujian@pku.edu.cn and tzliu@jlu.edu.cn

Received 24 August 2010
Accepted 21 November 2010
Published 23 December 2010

Online at stacks.iop.org/JSTAT/2010/P12030
doi:10.1088/1742-5468/2010/12/P12030

Abstract. One of the most relevant features of complex networks representing real systems is the community structure. In this paper, we extend the measure of diffusion distance between nodes in a network to a generalized form on the coarse-grained network with data parameterization via eigenmaps. This notion of proximity of meta-nodes in the coarse-grained networks reflects the intrinsic geometry of the partition in terms of connectivity of the communities in a diffusion process. Nodes are then grouped into communities through an agglomerative hierarchical clustering technique under this measure and the modularity function is used to select the best partition of the resulting dendrogram. The present algorithm can identify the community structure with a high degree of efficiency and accuracy. An appropriate number of communities can be automatically determined without any prior knowledge about the community structure. The computational results on several artificial and real-world networks confirm the capability of the algorithm.

Keywords: network dynamics

[3] Author to whom any correspondence should be addressed.

©2010 IOP Publishing Ltd and SISSA
1742-5468/10/P12030+24$30.00

Contents

1. Introduction **2**

2. The coarse-grained diffusion distance between communities in complex networks **4**
2.1. Framework of random walk and diffusion distance 4
2.2. Construction of coarse-grained random walk and coarse-grained diffusion distance 5

3. The algorithm **9**
3.1. Modularity maximization and its main limits 9
3.2. The agglomerative algorithm based on coarse-grained diffusion distance . . 10
3.3. Computational complexity 12

4. Experimental results **13**
4.1. Tests on artificial networks 13
4.1.1. Ad hoc networks. 13
4.1.2. The LFR benchmark. 14
4.2. Application to real-world networks 15
4.2.1. The karate club network. 15
4.2.2. The dolphins network. 17
4.2.3. The political books network. 18
4.2.4. The football team network. 20
4.2.5. The SFI collaboration network. 20

5. Conclusions **22**

Acknowledgments **23**

References **23**

1. Introduction

We have seen an explosive growth of interest and activity on the structure and dynamics of complex networks during recent years [1]–[3]. This is partly due to the influx of new ideas, particularly ideas from statistical mechanics, to the subject, and partly due to the emergence of interesting and challenging new examples of complex networks such as the Internet and wireless communication networks. Network models have also become popular tools in social science, economics, the design of transportation and communication systems, banking systems, powergrids, etc, due to our increased capability of analyzing the models [4]–[6]. On a related but different front, recent advances in computer vision and data mining have also relied heavily on the idea of viewing a data set or an image as a graph or a network, in order to extract information about the important features of the images or more generally, the data sets [7]–[9]. Since these networks are typically very complex, it is of great interest to see whether they can be reduced to much simpler systems. In particular, much effort has gone into partitioning the network into a small

doi:10.1088/1742-5468/2010/12/P12030

J. Stat. Mech. (2010) P12030

number of communities [10]–[40], which are constructed from different viewing angles in the literature. And in a broader aspect, it is also closely related to the model reduction theory of differential equations [41].

Random walks is among the most popular models in community structure detection [18]–[25]. If a network has a strong community structure, a random walker spends a long time inside a community due to the high density of internal edges and consequent number of paths that could be followed. In [19], random walks was used to define a distance between pairs of vertices: the distance between node x and y is the average number of edges that a random walker has to cross to reach y starting from x. This basic idea was extended by defining, based on this distance matrix, a quantity called the dissimilarity index between nearest-neighboring nodes [20]. The dissimilarity index signified to what extent two nearest-neighboring nodes would like to be in the same community. A different distance measure between nodes was introduced in [21], which is calculated from the probabilities that the random walker moves from a node to another in a fixed number of steps. In [22], the authors designed a graph clustering technique based on a signaling process between nodes, somewhat resembling diffusion. In [23] the authors made clusters via reducing the Markov chain. Network partitioning turns into a coding problem in that finding the partition yields the minimum description length of an infinite random walk. In the methods in [24, 25], the best partition of a graph is such that the Markov chain describing a random walk on the meta-network gives the best approximation of the full random walk dynamics on the whole network. The quality of the approximation is given by the distance between the transition matrices of the two processes, which thus needs to be minimized.

In the current paper, we extend the measure of diffusion distance [21] between nodes in a complex network to a generalized form on the coarse-grained network whose nodes are the communities of the original network, with data parameterization via eigenmaps. This notion of proximity of 'nodes' in the coarse-grained networks reflects the intrinsic geometry of the meta-node set in terms of connectivity of the communities in a diffusion process. The diffusion distance between two communities will be small if they are connected by many paths on the meta-network. This metric is thus a key quantity in the design of algorithm that are based on the preponderance of evidence for a given hypothesis. Suppose one wants to infer community labels for nodes based on a small number of labeled ones. This metric is then usually more appropriate than the linkage choices in traditional clustering literature [42], such as single linkage and complete linkage, as it takes into account all the information relating the two communities. Furthermore, since diffusion-based distances add up the contribution from all the possible paths, they are also robust to noise. Nodes are then grouped into communities through an agglomerative hierarchical clustering technique [42], and the modularity function [11]–[16], [21, 28, 36] is used to select the best partition of the resulting dendrogram.

We constructed our algorithm—agglomerative method based on the coarse-grained diffusion distance (AMCD) for network partition. The algorithm is tested on two artificial networks, including the ad hoc networks and the LFR benchmark networks. The algorithm is efficiently implemented with reasonable computational effort and leads to an accurate partitioning result. Moreover, successful application to several real-world networks, including the karate club network, the dolphins network, the American political

J. Stat. Mech. (2010) P12030

books network, the football team network and the SFI collaboration network, confirm the effectiveness of the present algorithm.

The rest of the paper is organized as follows. In section 2, we briefly introduce the framework of random walk and diffusion distance for proximity of nodes, then derive the measure of coarse-grained diffusion distance for proximity of communities. After reviewing the concept of modularity and its two main limits, we describe our algorithm in section 3. The advantages over linkage choices, and the computational complexity of the method, are also discussed in detail. In section 4, we apply the proposed algorithm to the representative examples mentioned before. Finally, we draw our conclusions in section 5.

2. The coarse-grained diffusion distance between communities in complex networks

2.1. Framework of random walk and diffusion distance

We will start with a brief review of random walks on complex networks [18]–[25]. Let $G(S, E)$ be a network with n nodes and m edges, where S is the set of nodes, $E = \{e(x, y)\}_{x,y\in S}$ is the weight matrix and $e(x, y)$ is the weight for the edge connecting the node x and y. The total weight of edges is denoted by m_e, and naturally $m_e = \sum_{x,y\in S} e(x, y)/2$. A simple example of the weight matrix is given by the adjacency matrix: $e(x, y) = 0$ or 1, depending on whether x and y are connected.

Let us consider a discrete random walk process, or diffusion process, on the network $G(S, E)$. At each time step, a walker is on a node and moves to a node chosen randomly and uniformly among its neighbors. The sequence of visited nodes is a Markov chain, the states of which are the nodes of the network. At each step, the transition probability from node x to node y is given by

$$p(x, y) = \frac{e(x, y)}{d(x)}, \qquad d(x) = \sum_{z\in S} e(x, z), \tag{1}$$

where $d(x)$ is the degree of the node x [43, 44]. This defines the transition matrix $P = \{p(x, y)\}_{x,y\in S}$ of random walk processes. By construction, this quantity reflects the first-order neighborhood structure of the network.

The basic idea introduced in the diffusion process framework [18] is to capture information on larger neighborhoods by taking powers of the matrix P, or equivalently, to run the random walk forward in time. The process is driven by the $P^t = \{p_t(x, y)\}_{x,y\in S}$, where $p_t(x, y)$ represents the probability of going from node x to node y through a random walk in t time steps, and $p_1(x, y) = p(x, y)$ for consistency. Increasing t, corresponds to propagating the local influence of each node with its neighbors. In other words, the quantity P^t reflects the intrinsic geometry of the node set S, defined via the connectivity of the network in a diffusion process, and the time t of the diffusion process plays the role of a scale parameter in the analysis.

When the time step t of a random walk starting at node x tends towards infinity, the probability of being on a node y only depends on the degree of y and not on the starting node x, that is

$$\lim_{t\to+\infty} p_t(x, y) = \mu(y), \qquad x, y \in S, \tag{2}$$

J. Stat. Mech. (2010) P12030

where μ is the stationary distribution of this Markov chain with the form

$$\mu(x) = \frac{d(x)}{\sum_{z\in S} d(z)}, \qquad x \in S. \tag{3}$$

Note that the stationary distribution $\mu(x)$ is, up to normalization, equal to the degree of the node x, $d(x)$, which means the more links a node has to other nodes in the network, the more often it will be visited by a random walker. Here we restrict ourselves to an undirected network, i.e. $e(x, y) = e(y, x)$, then μ satisfies the detailed balance condition

$$\mu(x)p(x, y) = \mu(y)p(y, x), \tag{4}$$

and the Markov chain is then reversible.

For a fixed but finite value $t > 0$, we want to define a metric between nodes in S which is such that two nodes x and y will be close if the corresponding conditional distributions $p_t(x, \cdot)$ and $p_t(y, \cdot)$ are close. As in [21], the diffusion distance $D_t(x, y)$ between x and y is defined as the weighted L^2 distance

$$D_t^2(x, y) = \sum_{z\in S} \frac{(p_t(x, z) - p_t(y, z))^2}{\mu(z)}, \tag{5}$$

where the weight $\mu(z)^{-1}$ penalizes discrepancies on domains of low degree more than those of high degree. This notion of proximity of nodes in the networks reflects the intrinsic geometry of the set S in terms of connectivity of them in a diffusion process. The diffusion distance between two nodes will be small if they are connected by many paths in the network. Here, the L^2 metric between the conditional distribution has the advantage that it allows one to relate distances to the spectral properties of the random walk and thereby connect Markov random walk learning on networks with data parameterization via eigenmaps [44].

2.2. Construction of coarse-grained random walk and coarse-grained diffusion distance

As mentioned, an advantage of the above definition of the diffusion distance (5) is the connection to the spectral theory of the random walk. As is well known, the transition matrix P has a set of left and right eigenvectors and a set of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}| \geq 0$

$$P\varphi_i = \lambda_i\varphi_i, \qquad \psi_i^T P = \lambda_i\psi_i^T, \qquad i = 0, 1, \ldots, n-1. \tag{6}$$

Note that $\psi_0 = \mu$, $\varphi_0 \equiv 1$ and $\psi_i^T\varphi_j = \delta_{ij}$. The left and right eigenvectors are related according to

$$\psi_i(x) = \varphi_i(x)\mu(x), \qquad x \in S, \quad i = 0, 1, \ldots, n-1. \tag{7}$$

The spectral decomposition of P^t is given by

$$p_t(x, y) = \sum_{i=0}^{n-1} \lambda_i^t\varphi_i(x)\psi(y) = \sum_{i=0}^{n-1} \lambda_i^t\varphi_i(x)\varphi(y)\mu(y), \tag{8}$$

J. Stat. Mech. (2010) P12030

which corresponds to a weighted principal component analysis of P^t [44]. Then the diffusion distance (5) can be reduced to

$$\begin{aligned} D_t^2(x,y) &= \sum_{z\in S}\frac{1}{\mu(z)}\left(\sum_{i=0}^{n-1}\lambda_i^t\varphi_i(x)\varphi_i(z)\mu(z)-\sum_{i=0}^{n-1}\lambda_i^t\varphi_i(y)\varphi_i(z)\mu(z)\right)^2 \\ &= \sum_{z\in S}\frac{1}{\mu(z)}\left(\sum_{i=0}^{n-1}\lambda_i^t(\varphi_i(x)-\varphi_i(y))\varphi_i(z)\mu(z)\right)^2 \\ &= \sum_{z\in S}\mu(z)\sum_{i=0}^{n-1}\lambda_i^t(\varphi_i(x)-\varphi_i(y))\varphi_i(z)\sum_{j=0}^{n-1}\lambda_j^t(\varphi_j(x)-\varphi_j(y))\varphi_j(z) \\ &= \sum_{i,j=0}^{n-1}\lambda_i^t\lambda_j^t(\varphi_i(x)-\varphi_i(y))(\varphi_j(x)-\varphi_j(y))\sum_{z\in S}\mu(z)\varphi_i(z)\varphi_j(z) \\ &= \sum_{i=0}^{n-1}\lambda_i^{2t}(\varphi_i(x)-\varphi_i(y))^2. \end{aligned} \tag{9}$$

We take a partition of S as $S=\bigcup_{k=1}^{N}S_k$ with $S_k\bigcap S_l=\emptyset$ if $k\neq l$. Our aim is to aggregate the nodes in each community in order to coarse-grain both the state set S and the time evolution of the random walk. To do so, we regard each set S_k in the state space $\mathbb{S}=\{S_1,\ldots,S_N\}$ as corresponding to the nodes of a N-nodes network $\hat{G}(\mathbb{S},\mathbb{E}_t)$, where $\mathbb{E}_t=\{\hat{e}_t(S_k,S_l)\}_{S_k,S_l\in\mathbb{S}}$, and the weight $\hat{e}_t(S_k,S_l)$ on the edge that connects S_k and S_l is defined as

$$\hat{e}_t(S_k,S_l)=\sum_{x\in S_k,y\in S_l}\mu(x)p_t(x,y), \tag{10}$$

where the sum involves all the transition probabilities between $x\in S_k$ and $y\in S_l$. From the detailed balance condition (4), it can be verified that $\hat{e}_t(S_k,S_l)=\hat{e}_t(S_l,S_k)$. By setting

$$\hat{\mu}(S_k)=\sum_{z\in S_k}\mu(z),\qquad k=1,\ldots,N, \tag{11}$$

one can define a coarse-grained Markov chain on graph $\hat{G}(\mathbb{S},\mathbb{E}_t)$ with stationary distribution $\hat{\mu}$ and transition probabilities

$$\hat{p}_t(S_k,S_l)=\frac{\hat{e}_t(S_k,S_l)}{\sum_{m=1}^{N}\hat{e}_t(S_k,S_m)}=\frac{1}{\hat{\mu}(S_k)}\sum_{x\in S_k,y\in S_l}\mu(x)p_t(x,y). \tag{12}$$

It can be easily shown that $\hat{p}_t$ is a stochastic matrix on the state space $\mathbb{S}$ and satisfies a detailed balance condition with respect to $\hat{\mu}$, i.e.

$$\hat{\mu}(S_k)\hat{p}_t(S_k,S_l)=\hat{\mu}(S_l)\hat{p}_t(S_l,S_k). \tag{13}$$

This construction of a coarse-grained random walk is shown in figure 1.

Another approach of deriving the coarse-grained transition probability matrix (12) is described in [24, 25]. Any coarse-grained stochastic matrix $\hat{p}_t$ can be naturally lifted to

J. Stat. Mech. (2010) P12030

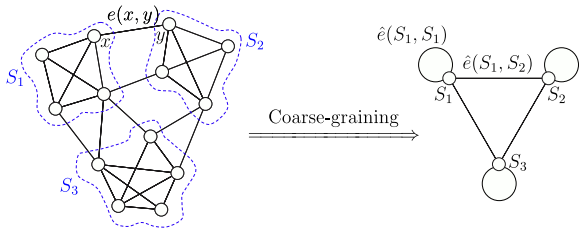

Figure 1. Coarse-grained random walk of a network. For a given partition $S = S_1 \bigcup S_2 \bigcup S_3$ on a network $G(S, E)$, we define a coarse-grained network $\hat{G}(\mathbb{S}, \mathbb{E}_t)$ by aggregating all nodes belonging to a subset S_k into a meta-node. The new weights $\hat{e}(S_k, S_l)$ are computed via weight averaging the transition probabilities between nodes $x \in S_k$ and $y \in S_l$, $k, l = 1$–3, and a new Markov chain with transition probabilities $\hat{p}(S_k, S_l)$ can also be obtained.

the space of stochastic matrices on the original state space S via

$$\tilde{p}_t(x, y) = \sum_{k,l=1}^{N} \mathbf{1}_{S_k}(x)\mathbf{1}_{S_l}(y)\hat{p}_t(S_k, S_l)\frac{\mu(y)}{\hat{\mu}(S_l)}, \tag{14}$$

where $\mathbf{1}_{S_k}(x) = 1$ if $x \in S_k$ and $\mathbf{1}_{S_k}(x) = 0$ otherwise. The basic idea in [24, 25] is to introduce a metric, also called the Hilbert–Schmidt norm, in the space of stochastic matrices. The optimal partition and the corresponding reduced Markov chain $\tilde{p}_t$ is found by minimizing

$$\|p_t - \tilde{p}_t\|_\mu^2 = \sum_{x,y \in S} \frac{\mu(x)}{\mu(y)}(p_t(x, y) - \tilde{p}_t(x, y))^2. \tag{15}$$

Based upon this formulation, we can find the optimal $\hat{p}_t(S_k, S_l)$ for any fixed partition $\{S_1, \ldots, S_N\}$. From the optimality condition

$$\begin{aligned}
\frac{\partial \|p_t - \tilde{p}_t\|_\mu^2}{\partial \hat{p}_t(S_k, S_l)} &= 2\sum_{x,y \in S} \frac{\mu(x)}{\mu(y)}(p_t(x, y) - \tilde{p}_t(x, y))\frac{\partial \tilde{p}_t(x, y)}{\partial \hat{p}_t(S_k, S_l)} \\
&= 2\sum_{x,y \in S} \frac{\mu(x)}{\mu(y)} \left(p_t(x, y) - \sum_{m,n=1}^{N} \mathbf{1}_{S_m}(x)\mathbf{1}_{S_n}(y)\hat{p}_t(S_m, S_n)\frac{\mu(y)}{\hat{\mu}(S_n)} \right) \\
&\quad \cdot \mathbf{1}_{S_k}(x)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)} \\
&= 2\sum_{x \in S_k, y \in S_l} \frac{\mu(x)}{\mu(y)} \left(p_t(x, y) - \hat{p}_t(S_k, S_l)\frac{\mu(y)}{\hat{\mu}(S_l)} \right) \frac{\mu(y)}{\hat{\mu}(S_l)} \\
&= \frac{2}{\hat{\mu}(S_l)} \left(\sum_{x \in S_k, y \in S_l} \mu(x)p_t(x, y) - \hat{p}_t(S_k, S_l)\hat{\mu}(S_k) \right) = 0,
\end{aligned} \tag{16}$$

we can obtain (12). It indicates that when the partition $\{S_1, \ldots, S_N\}$ is known, the minimizer of (15) is unique, which can be given by (12), and the corresponding $\tilde{p}_t$ is the

J. Stat. Mech. (2010) P12030

stochastic matrix in the class (14) which provides the best rank-N approximation of the original one under the metric (15).

More generally, we define coarse-grained versions of ψ_i in a similar way by summing over the nodes in a partition

$$\hat{\psi}_i(S_k) = \sum_{z \in S_k} \psi_i(z), \qquad k = 1, \ldots, N, \tag{17}$$

and as in (7), coarse-grained versions of φ_i, according to the duality condition $\hat{\psi}_i(S_k) = \hat{\varphi}_i(S_k)\hat{\mu}(S_k)$, are defined as

$$\hat{\varphi}_i(S_k) = \frac{\hat{\psi}_i(S_k)}{\hat{\mu}(S_k)} = \frac{1}{\hat{\mu}(S_k)} \sum_{z \in S_k} \varphi_i(z)\mu(z), \qquad k = 1, \ldots, N. \tag{18}$$

Then the coarse-grained probability $\hat{p}_t$ can be written in a similar way to (8) in the spectral decomposition form as follows

$$\begin{aligned} \hat{p}_t(S_k, S_l) &= \frac{1}{\hat{\mu}(S_k)} \sum_{x \in S_k, y \in S_l} \mu(x) \sum_{i=0}^{n-1} \lambda_i^t \varphi_i(x)\psi_i(y) \\ &= \sum_{i=0}^{n-1} \lambda_i^t \cdot \frac{1}{\hat{\mu}(S_k)} \sum_{x \in S_k} \varphi_i(x)\mu(x) \cdot \sum_{y \in S_l} \psi_i(y) \\ &= \sum_{i=0}^{n-1} \lambda_i^t \hat{\varphi}_i(S_k)\hat{\psi}_i(S_l) = \sum_{i=0}^{n-1} \lambda_i^t \hat{\varphi}_i(S_k)\hat{\varphi}_i(S_l)\hat{\mu}(S_l). \end{aligned} \tag{19}$$

This can be considered as an extension version of (8). For a fixed but finite value $t > 0$, we want to define a metric between meta-nodes in $\mathbb{S}$ which is such that two communities S_k and S_l will be close if the corresponding conditional distributions $\hat{p}_t(S_k, \cdot)$ and $\hat{p}_t(S_l, \cdot)$ are close. As mentioned above, the diffusion distance between community S_k and S_l is given by

$$\begin{aligned} \hat{D}_t^2(S_k, S_l) &= \sum_{m=1}^{N} \frac{(\hat{p}_t(S_k, S_m) - \hat{p}_t(S_l, S_m))^2}{\hat{\mu}(S_m)} \\ &= \sum_{i,j=0}^{n-1} \lambda_i^t \lambda_j^t (\hat{\varphi}_i(S_k) - \hat{\varphi}_i(S_l))(\hat{\varphi}_j(S_k) - \hat{\varphi}_j(S_l)) \sum_{m=1}^{N} \hat{\psi}_i(S_m)\hat{\varphi}_j(S_m). \end{aligned} \tag{20}$$

Here, the term $\psi_i^T \varphi_j$ no longer equals the Kronecker delta function. This notion of proximity of communities in the coarse-grained networks reflects the intrinsic geometry of the set $\mathbb{S}$ in terms of connectivity of the meta-nodes in a diffusion process. The diffusion distance between S_k and S_l will be small if they are connected by many paths in the meta-network. This metric is thus a key quantity in the design of the following algorithm, which is based on the preponderance of evidence for a given hypothesis. For example, suppose one wants to infer community labels for nodes in $\mathbb{S}$ based on a small number of labeled examples. Then one can easily propagate the label information from a labeled example S_k to the new node S_l following (i) the shortest path, or (ii) all paths connecting

J. Stat. Mech. (2010) P12030

S_k to S_l. The latter one is usually more appropriate, as it takes into account all evidence relating S_k to S_l. Furthermore, since diffusion distances add up the contribution from all the possible paths, they are also robust to noise.

3. The algorithm

3.1. Modularity maximization and its main limits

In recent years, a concept of modularity proposed by Newman [11]–[16], [21, 28, 36] has been widely used as a measure of goodness for community structure. A good division of a network into communities is not merely one in which the number of edges running between groups is small. Rather, it is one in which the number of edges between groups is smaller than expected. These considerations lead to the modularity Q defined by

$Q =$ (number of edges within communities)
$-$ (expected number of such edges).

It is a function of the particular partition of the network into groups, with larger values indicating stronger community structure.

The definition of the modularity can involve a comparison of the number of within-group edges in a real network and the number in some equivalent randomized model network in which edges are placed without regard to community structure [15]. The null model also has n nodes as the original network. The probability $p^{\mathrm{E}}(x, y)$ for an edge to fall between every pair of node x and y is specified. More precisely, $p^{\mathrm{E}}(x, y)$ is the expected number of edges between x and y, a definition that allows for the possibility that there may be more than one edge between a pair of nodes, which happens in certain types of networks. For a given partition $\{S_k\}_{k=1}^{N}$, the modularity can be written as

$$Q = \frac{1}{2m_{\mathrm{e}}} \sum_{k=1}^{N} \sum_{x,y \in S_k} (e(x, y) - p^{\mathrm{E}}(x, y)), \qquad p^{\mathrm{E}}(x, y) = \frac{d(x)d(y)}{2m_{\mathrm{e}}} \tag{21}$$

and m_{e} is the total weight of edges given by $\sum_{x,y \in S} e(x, y)/2$. This model is closely related to the configuration model, which has been widely studied in physics [15, 44]. Some existing methods are presented to find good partitions of a network into communities by optimizing the modularity over possible divisions, which has proven highly effective in practice [11]–[16], [28, 37, 39].

Despite the popularity of modularity maximization, much remains unknown about the quality and significance of its output when applied to real-world networks with unknown community structure. Modularity is not an exact indicator of community structure for two main reasons. The first is that this quantity has a well known resolution limit phenomenon that makes its applicability questionable to large graphs, as small modules remain undetectable [45]. Modularity optimization may fail to identify modules smaller than a scale which depends on the total weight of edges m_{e} of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined. The probability that a module conceals well-defined substructures is the highest if the number of links internal to the module is of the order of $\sqrt{2m_{\mathrm{e}}}$ or smaller. It is thus *a priori* impossible to tell whether a module obtained through modularity optimization is

J. Stat. Mech. (2010) P12030

indeed a single module or a cluster of smaller modules. This result thus introduces some caveats in the use of modularity to detect community structure.

Secondly, the modularity landscape is 'glassy', i.e. it is made by a huge number of partitions with modularity very close to the maximum, but structurally different, so the meaning of the maximum is also under debate [46]. Modularity exhibits extreme degeneracies: it typically admits an exponential number of distinct high scoring solutions and typically lacks a clear global maximum. The degenerate solutions can fundamentally disagree on many, but not all, partition properties, such as the composition of the largest modules and the distribution of module sizes. These results imply that the output of any modularity maximization procedure should be interpreted cautiously.

The modularity itself has however not yet been thoroughly investigated and only a few general properties are known. As mentioned in [45], quality functions are still helpful, but their role should be probably limited to the comparison of partitions with the same number of modules.

3.2. The agglomerative algorithm based on coarse-grained diffusion distance

Agglomerative clustering algorithms begin with every observation representing a singleton cluster. At each of the $n-1$ steps the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level [42]. Here we make use of the agglomerative hierarchical clustering technique for network partition and choose the coarse-grained diffusion distance as the measure of dissimilarity between communities at each step. The maximal value of modularity determines the optimal partition of the network. The whole procedure for the agglomerative method based on coarse-grained diffusion distance (AMCD) is summarized below.

(1) Set parameter t and initialize the partition $\mathbb{S}^{(1)} = \{S_1^{(1)}, \ldots, S_{N^{(1)}}^{(1)}\}$, $N^{(1)} = n$, with a single node in each community, and calculate $Q^{(1)}$ according to $\mathbb{S}^{(1)}$; set $r^* = 1$.

(2) Compute λ_i, φ_i, ψ_i, $i = 0, 1, \ldots, n-1$ and the diffusion distance matrix D_t. Here we only have to store its strictly upper triangular matrix form due to its symmetry; set $\hat{D}_t^{(1)} = D_t$.

(3) For $r = 1, 2, \ldots, n-1$, do the following

  (3.1) Find the minimal entry $\hat{D}_t^{(r)}(S_{k^{(r)}}^{(r)}, S_{l^{(r)}}^{(r)})$ in the strictly upper triangular matrix $\hat{D}_t^{(r)}$, where $S_{k^{(r)}}^{(r)}, S_{l^{(r)}}^{(r)} \in \mathbb{S}^{(r)} = \{S_1^{(r)}, \ldots, S_{N^{(r)}}^{(r)}\}$, $1 \le k^{(r)} < l^{(r)} \le N^{(r)}$ are the two communities that need to be aggregated together.

  (3.2) Update the partition $\mathbb{S}^{(r+1)} = \{S_1^{(r+1)}, \ldots, S_{n-r}^{(r+1)}\}$ by setting $S_{k^{(r)}}^{(r+1)} = S_{k^{(r)}}^{(r)} \bigcup S_{l^{(r)}}^{(r)}$, $S_{l^{(r)}+m-1}^{(r+1)} = S_{l^{(r)}+m}^{(r)}, m = 1, \ldots, N^{(r)} - l^{(r)}$ and keeping the other communities; set $N^{(r+1)} = n - r$.

  (3.3) Compute the current modularity $Q^{(r+1)}$ according to $\mathbb{S}^{(r+1)}$. Update the optimal state, i.e. if $Q^{(r+1)} > Q^{(r^*)}$, set $r^* = r + 1$.

  (3.4) Update the coarse-grained diffusion distance matrix $\hat{D}_t^{(r+1)}$.

(4) Output the optimal partition $\mathbb{S}^{(r^*)} = \{S_1^{(r^*)}, \ldots, S_{N^{(r^*)}}^{(r^*)}\}$ and the maximum $Q^{(r^*)}$ of the whole procedure.

J. Stat. Mech. (2010) P12030

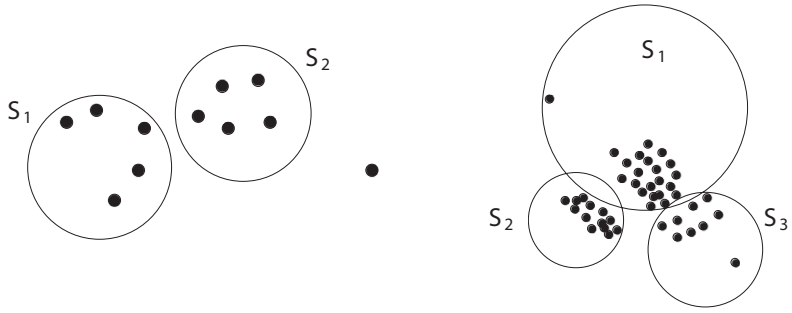

Figure 2. Problem with single linkage (left plot) and complete linkage (right plot).

Given a distance measure between points, the user has many choices for how to define intergroup similarity in traditional clustering literature [42]. However, different choices have different benefits and shortages. Now suppose we already have two communities S_k and S_l. The three most popular choices and their defects can be summarized as follows.

(i) Single linkage: the similarity of the closest pair

$$\hat{D}_t^{\mathrm{SL}}(S_k, S_l) = \min_{x \in S_k, y \in S_l} D_t(x, y). \tag{22}$$

This is also called the nearest-neighbor technique. Single linkage can produce chaining, where a sequence of close observations in different groups cause early merges of those groups. For example, in figure 2, suppose that the earlier grouping groups the two circled parts. The next grouping will group the two grouped parts and the individual point will be left alone.

(ii) Complete linkage: the similarity of the furthest pair

$$\hat{D}_t^{\mathrm{CL}}(S_k, S_l) = \max_{x \in S_k, y \in S_l} D_t(x, y). \tag{23}$$

Complete linkage has the opposite problem. It might not merge close groups because of outlier members that are far apart. For example, in figure 2, although S_1 and S_3 should have been clustered, with complete linkage, S_1 and S_2 are actually clustered.

(iii) Average linkage: the average similarity between groups

$$\hat{D}_t^{\mathrm{AL}}(S_k, S_l) = \frac{1}{|S_k||S_l|} \sum_{x \in S_k, y \in S_l} D_t(x, y), \tag{24}$$

where $|S_k|$ and $|S_l|$ are the respective number of observations in each group. Group average represents a natural compromise, but depends on the scale of the similarities. Applying a monotone transformation to the similarities can change the results.

The advantage of our algorithm is that the proposed measure of coarse-grained diffusion distance overcomes the weaknesses of the traditional linkage techniques based on node-to-node dissimilarity mentioned above, since it takes into account all the information relating the two clusters. Furthermore, since diffusion-based distances add up the contribution from all the possible paths, they are also robust to noise. Another advantage is that it

J. Stat. Mech. (2010) P12030

seems more efficient than the family of the k-means iterative methods [9, 24, 25, 36], which have to be implemented many times due to their local minima with iteration steps in each trial that are difficult to estimate. The only parameter in our computation is the time step t, and increasing t corresponds to propagating the local influence of each node with its neighbors. The agglomerative process can efficiently and automatically determine the number of communities N without fixing it as a known model parameter. Hence, the present algorithm can avoid ineffective repetition and lead to a high degree of efficiency and accuracy.

3.3. Computational complexity

The spectral approach takes an important part in the search for community structure in networks [15, 16]. However all these approaches have the same drawback that the eigenvectors need to be explicitly computed in time $O(n^3)$ for a sparse matrix. Our approach is based on the same foundation, but has the advantage of avoiding the expensive computation of the eigenvectors: it only needs to compute the t-step probabilities $P^t = \{p_t(x, y)\}_{x,y\in S}$, which can be done efficiently, as shown in [21]. To compute the probability vector $p_t(x, \cdot)$, we multiply t times the vector $p_0(x, \cdot)$ by the matrix P, here $p_0(x, z) = \delta(x, z)$ for all $z \in S$. This direct method is advantageous in our case because the matrix P is generally sparse for most real-world complex networks, therefore each product is processed in time $O(m)$ and m represents the number of edges, which is usually assumed to be $O(n)$ in such realistic networks. The initialization of $p_0(x, \cdot)$ is done in $O(n)$ and thus each of the n vectors $p_t(x, \cdot)$ is computed in time $O(n+tm) = O(tm)$. Consequently, the initialization of the probability vectors $\{p_t(x, \cdot)\}_{x\in S}$ is done in $O(mnt)$ and the item $\{\mu(x)p_t(x, \cdot)\}_{x\in S}$ costs $O(n^2)$.

At each step r of the algorithm, the computation for $\hat{p}_t$ costs $O(N)$ and the cost of computing $\hat{D}_t$ in step (3.4) is $O(N^2)$. Suppose the two communities that need to be aggregated are $S^{(r)}_{k^{(r)}}$ and $S^{(r)}_{l^{(r)}}$ and the new combined cluster is denoted by $S^{(r+1)}_{k^{(r)}}$, then we update $D^{(r+1)}_t(S_k, S_l)$ by considering the following two cases: if $k = k^{(r)}$ or $l = k^{(r)}$, we keep the computation in (20)

$$D^{(r+1)}_t(S_k, S_l)^2 = \sum_{m=1}^{N} \frac{(\hat{p}_t(S_k, S_m) - \hat{p}_t(S_l, S_m))^2}{\hat{\mu}(S_m)}, \tag{25}$$

otherwise we take the form

$$\begin{aligned} D^{(r+1)}_t(S_k, S_l)^2 = D^{(r)}_t(S_k, S_l)^2 &- \frac{(\hat{p}_t(S_k, S^{(r)}_{k^{(r)}}) - \hat{p}_t(S_l, S^{(r)}_{k^{(r)}}))^2}{\hat{\mu}(S^{(r)}_{k^{(r)}})} \\ &- \frac{(\hat{p}_t(S_k, S^{(r)}_{l^{(r)}}) - \hat{p}_t(S_l, S^{(r)}_{l^{(r)}}))^2}{\hat{\mu}(S^{(r)}_{l^{(r)}})} + \frac{(\hat{p}_t(S_k, S^{(r+1)}_{k^{(r)}}) - \hat{p}_t(S_l, S^{(r+1)}_{k^{(r)}}))^2}{\hat{\mu}(S^{(r+1)}_{k^{(r)}})}, \end{aligned} \tag{26}$$

and these two rules cost $O(N)$ and $O(N^2)$, respectively. So the total cost in the step of r is $O(N^2)$, where $N = n - r, r = 0, 1, \ldots, n - 1$. Therefore, the computational complexity of our method is $O((n^3/6) + mnt) = O(n^3/6)$. This computation rapidly becomes untractable in practice when the size of the network exceeds some thousands of vertices. How to give an approximation, or make use of other approaches to reduce

J. Stat. Mech. (2010) P12030

Table 1. The optimal number of communities and corresponding modularity obtained by our method based on the measure of coarse-grained diffusion distance (CD), compared with single linkage (SL), complete linkage (CL) and average linkage (AL), for the five real-world networks.

	Karate club ($t = 7$)		Dolphins ($t = 10$)		Political books ($t = 5$)		Football team ($t = 1$)		SFI collaboration ($t = 9$)	
	N	Q	N	Q	N	Q	N	Q	N	Q
CD	3	0.3991	2	0.3787	5	0.5202	11	0.6042	6	0.7266
SL	8	0.1206	2	0.3787	5	0.4903	14	0.5725	9	0.7021
CL	3	0.3991	7	0.4292	4	0.5262	15	0.5876	9	0.7103
AL	3	0.3744	3	0.4341	5	0.4842	6	0.4006	5	0.7079

computational cost, and further test our method on modern large datasets will be our next step. But the measure of coarse-grained diffusion distance considered in this paper is efficient, especially in overcoming the shortages of the traditional diffusion distance with linkage choices in community detection (see table 1 and figure 6), and deserved to be investigated.

4. Experimental results

4.1. Tests on artificial networks

In this subsection, we test our algorithm on two classical artificial networks with a known community structure, including the ad hoc networks and the LFR benchmark networks.

4.1.1. Ad hoc networks. The first example is the ad hoc network with 128 nodes. The ad hoc network is a benchmark problem used in many papers [10, 11, 13, 20, 22, 24, 25, 27, 28], [36]–[39], It has a known community structure and is constructed as follows. Suppose we choose $n = 128$ nodes and split them into four communities with 32 nodes each. Assume that pairs of nodes belonging to the same communities are linked with probability p_{in} and pairs belonging to different communities with probability p_{out}. These values are chosen so that the average node degree $\langle d \rangle$ is fixed at $\langle d \rangle = 16$. In other words, p_{in} and p_{out} are related as

$$31p_{\text{in}} + 96p_{\text{out}} = 16. \tag{27}$$

We will denote $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. To compare the built-in modular structure with the one delivered by the algorithm, we adopt the fraction of correctly identified nodes and the normalized mutual information, which have proved to be reliable [37]–[39]. It is based on defining a confusion matrix M, where the rows correspond to the real communities, and the columns correspond to the found communities. The member of M, M_{kl} is simply the number of nodes in the real community S_k that appear in the found community S_l. The number of real communities is denoted N_{r} and the number of found communities is denoted N_{f}, the sum over row k of matrix M_{kl} is denoted M_k and the sum over column l is denoted M_l. A measure of similarity

J. Stat. Mech. (2010) P12030

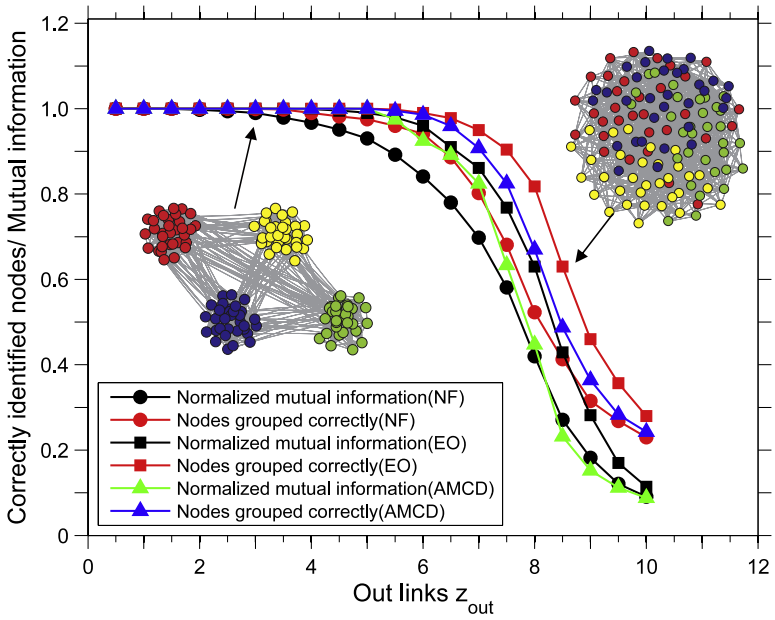

Figure 3. Test of our algorithm compared with Newman's fast algorithm [13] and the extremal optimization algorithm [28] on the ad hoc network with 128 nodes. The ad hoc network has four communities: for lower $z_{\rm out}$ the communities are easily distinguished, while for higher $z_{\rm out}$ this becomes more complicated.

between the partitions, based on information theory, is then

$$I(\mathbb{S}_{\rm r},\mathbb{S}_{\rm f})=\frac{-2\sum_{k=1}^{N_{\rm r}}\sum_{l=1}^{N_{\rm f}}M_{kl}\log(nM_{kl}/M_kM_l)}{\sum_{k=1}^{N_{\rm r}}M_k\log(M_k/n)+\sum_{l=1}^{N_{\rm f}}M_l\log(M_l/n)}. \tag{28}$$

We change $z_{\rm out}=96p_{\rm out}$ from 0.5 to 10 and look into the fraction of correctly identified nodes and the normalized mutual information. As we can see from figure 3, both two measures vary in a similar way across the different methods as $z_{\rm out}$ increases and the communities become more diffuse at the same time. The natural partition is always found up until $z_{\rm out}=6$, then the method starts to fail. It seems that our algorithm performs competitively with Newman's fast algorithm [13] and the extremal optimization algorithm [28], especially for the more complicated cases when $z_{\rm out}$ is higher.

4.1.2. The LFR benchmark. The LFR benchmark [38, 39] is a realistic network for community detection that accounts for the heterogeneity of both degree and community size. The node degrees are distributed according to a power law with exponent γ and the community sizes also obey a power law distribution with exponent β. In the construction of the benchmark networks, each node receives its degree once and for all and keeps it fixed until the end. It is more practical to choose as independent parameter the mixing parameter μ, which expresses the ratio between the external degree of a node with respect to its community and the total degree of the node.

In figure 4, we show what happens if our algorithm is implemented on the benchmark with $n=500$. The four panels correspond to four pairs for the exponents $(\gamma,\beta)=(2,\ 1)$, $(2,\ 2)$, $(3,\ 1)$, $(3,\ 2)$. We have chosen combinations of the extremes of the exponents'

J. Stat. Mech. (2010) P12030

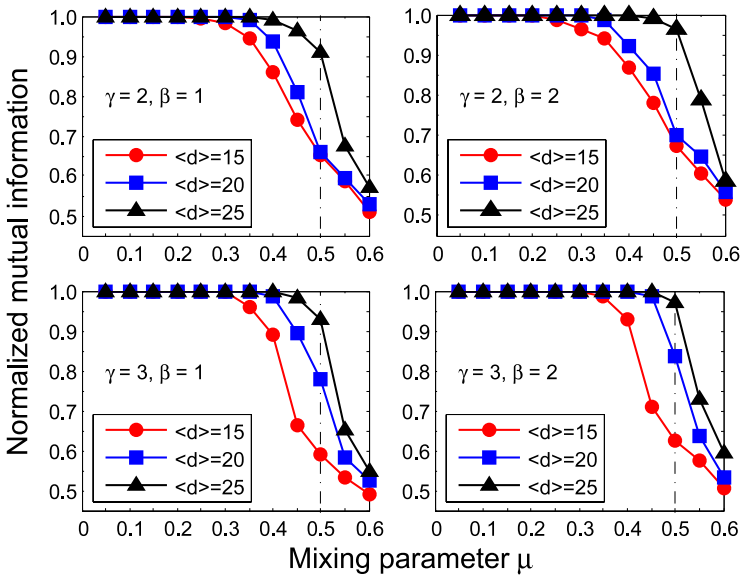

Figure 4. Test of our algorithm on the LFR benchmark [38, 39]. The number of nodes $n = 500$. The results clearly depend on all parameters of the benchmark, from the exponents γ and β to the average degree $\langle d \rangle$. The threshold $\mu_c = 0.5$, shown by the dashed vertical line in the plots, marks the border beyond which communities are no longer defined in the strong sense, i.e., such that each node has more neighbors in its own community than in the others.

ranges in order to explore the widest spectrum of network structures. Each curve shows the variation of the normalized mutual information with the mixing parameter μ. We can see that the performance of the method is better the larger the average degree $\langle d \rangle$, whereas it gets worse when the mixing parameter became larger. The threshold $\mu_c = 0.5$ shown by the dashed vertical line in the plots, marks the border beyond which communities are no longer defined in the strong sense, i.e., such that each node has more neighbors in its own community than in the others. In general, we can infer that our method gives good results.

4.2. Application to real-world networks

In this subsection, we describe the applications of our algorithm to five further real-world networks, including the karate club network, the dolphins network, the American political books network, the football team network and the SFI collaboration network.

4.2.1. The karate club network. This network was constructed by Zachary after he observed social interactions between members of a karate club at an American university [47]. Soon after, a dispute arose between the club administrator and the main teacher, and the club split into two smaller clubs. It has been used widely to test algorithms for finding communities in networks [10, 11, 13, 14, 16, 17], [19]–[22], [24]–[26], [28, 34, 36]. The modularity function Q change with the number of communities N in each agglomerative step for different time parameter t is shown in figure 5(a). The parameter is set by $t = 7$ in

J. Stat. Mech. (2010) P12030

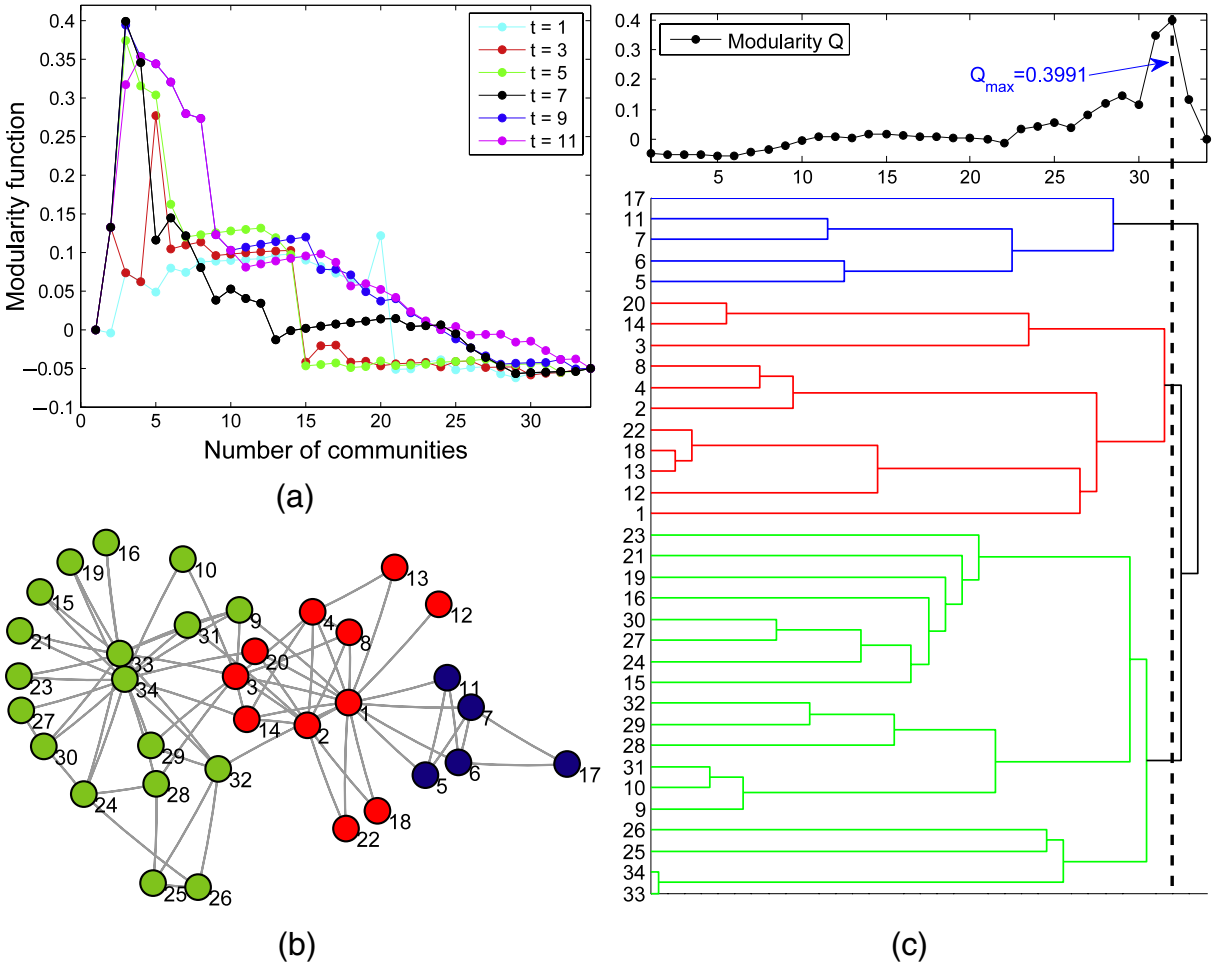

Figure 5. The computational results for the karate club network detected by our method. (a) The modularity change with number of communities in each iteration for different time parameter t. (b) The community structure identified by setting $t = 7$ corresponds to three communities represented by the colors. (c) The dendrogram of hierarchical structures and the optimal partition with a maximal modularity $Q = 0.3991$ is denoted by a vertical dashed line.

this model computation and the community structure detected by our method, shown in figure 5(b), corresponds to three communities represented by the colors. In figure 5(c) we give the dendrogram of the hierarchical structures. The optimal partition with a maximal modularity $Q = 0.3991$ is denoted by a vertical dashed line.

Now let us compare the community structure obtained by our method with the original partition result obtained by Zachary. In [47], Zachary gave the partition $S_1 = \{1 : 8, 11 : 14, 17, 18, 20, 22\}$ and $S_2 = \{9, 10, 15, 16, 19, 21, 23 : 34\}$. As shown in figure 5(b), almost all the nodes in S_2 are correctly clustered, and S_1 seems to be split into two clusters colored red and blue, which is also in accord with the network topology.

The final number of communities N and the maximum value of Q, comparing against the agglomerative method under the traditional diffusion distance and the three linkage choices, respectively, are presented in table 1. It seems that our algorithm performs better

J. Stat. Mech. (2010) P12030

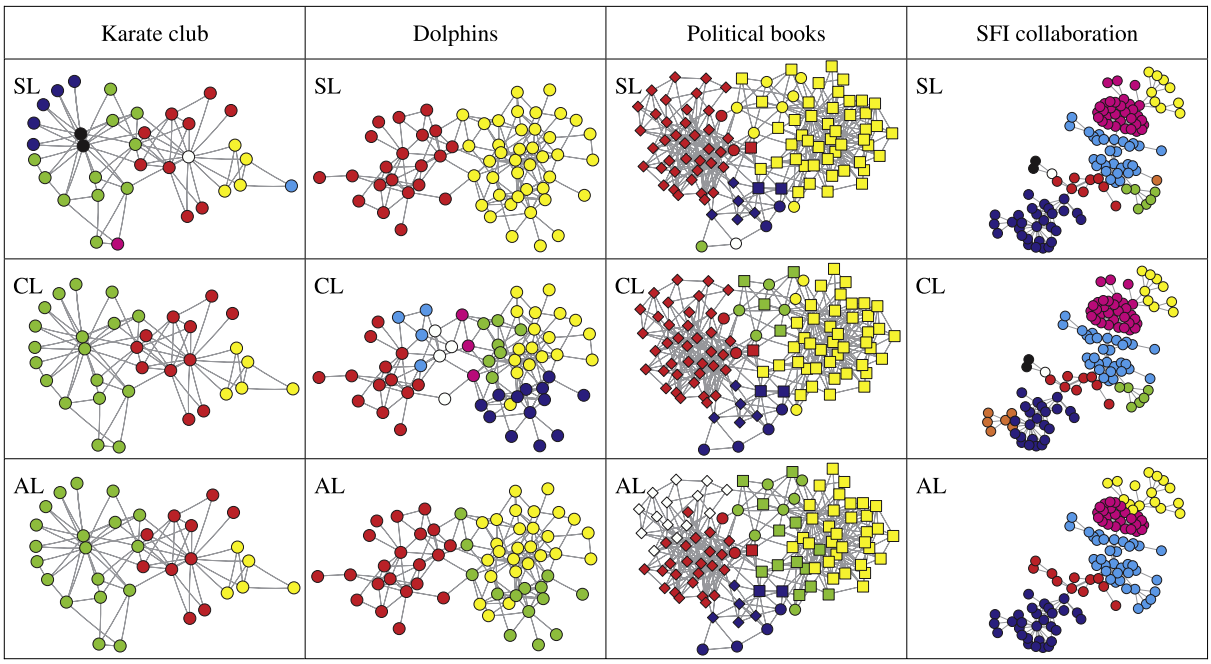

Figure 6. The partitioning results detected by single linkage (SL), complete linkage (CL) and average linkage (AL) for the real-world networks.

than SL and AL and produces the same result as CL. The communities of these networks detected by the three linkage choices are shown in figure 1, which embodies one of our conclusions that the proposed measure can always produce more appropriate clustering results than traditional linkage choices [42].

4.2.2. The dolphins network. The dolphins network is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand [48, 49]. The network was compiled from the studies of the dolphins, with ties between dolphin pairs being established by observation of statistically significant frequent association [11], [13]–[15], [36]. The modularity function Q change with the number of communities N in each agglomerative step for different time parameter t is shown in figure 7(a). The parameter is set by $t = 10$ in this model computation, and the community structure detected by our method, shown in figure 7(b), corresponds to two communities represented by the colors. In figure 7(c), we give the dendrogram of the hierarchical structures. The optimal partition with a maximal modularity $Q = 0.3787$ is denoted by a vertical dashed line. According to the results, the network seems to split into two large communities: the red part and the cyan part. This split into two groups appears to correspond to a known division of the dolphin community [49]. Actually, these dolphins separated in two groups after a dolphin left the place for some time. Such groups are quite cohesive, with several internal cliques, and easily identifiable: only six edges join nodes of different communities. Due to this natural classification, the dolphins network, like the karate club network, is often used to test algorithms for community detection. We list the optimal number of communities N and corresponding Q in table 1, comparing

J. Stat. Mech. (2010) P12030

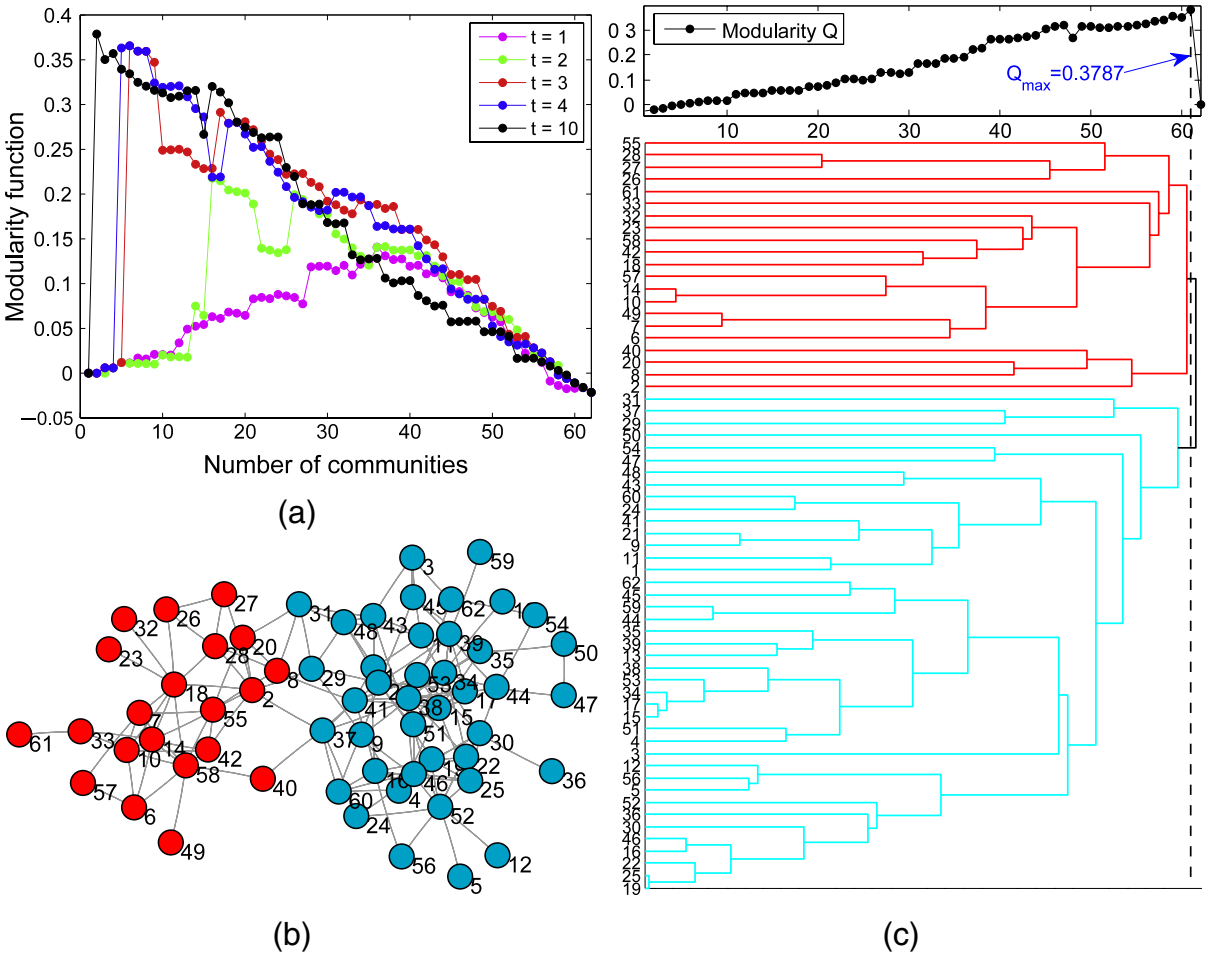

Figure 7. The computational results for the dolphins network detected by our method. (a) The modularity change with number of communities in each iteration for different time parameter t. (b) The community structure identified by setting $t = 10$ corresponds to two communities represented by the colors. (c) The dendrogram of the hierarchical structures and the optimal partition with a maximal modularity $Q = 0.3787$ is denoted by a vertical dashed line.

them against the agglomerative method based on the measure of node-to-node diffusion distance and associated with the three linkage choices, respectively. It seems that our algorithm produces the same result as CL. From the clustering results detected by linkage choices shown in figure 1, we can find that our method performs better than CL and AL, even if they can reach a higher value of modularity.

4.2.3. The political books network. We consider the network of books on politics, which are assigned based on a reading of the descriptions and reviews of the books posted on Amazon [15, 16]. In this network, the nodes represent 105 recent books on American politics bought from the on-line bookseller Amazon.com, and the edges join pairs of books that are frequently purchased by the same buyer, as indicated by the feature that customers who bought this book also bought these other books. As shown in figure 8(b),

J. Stat. Mech. (2010) P12030

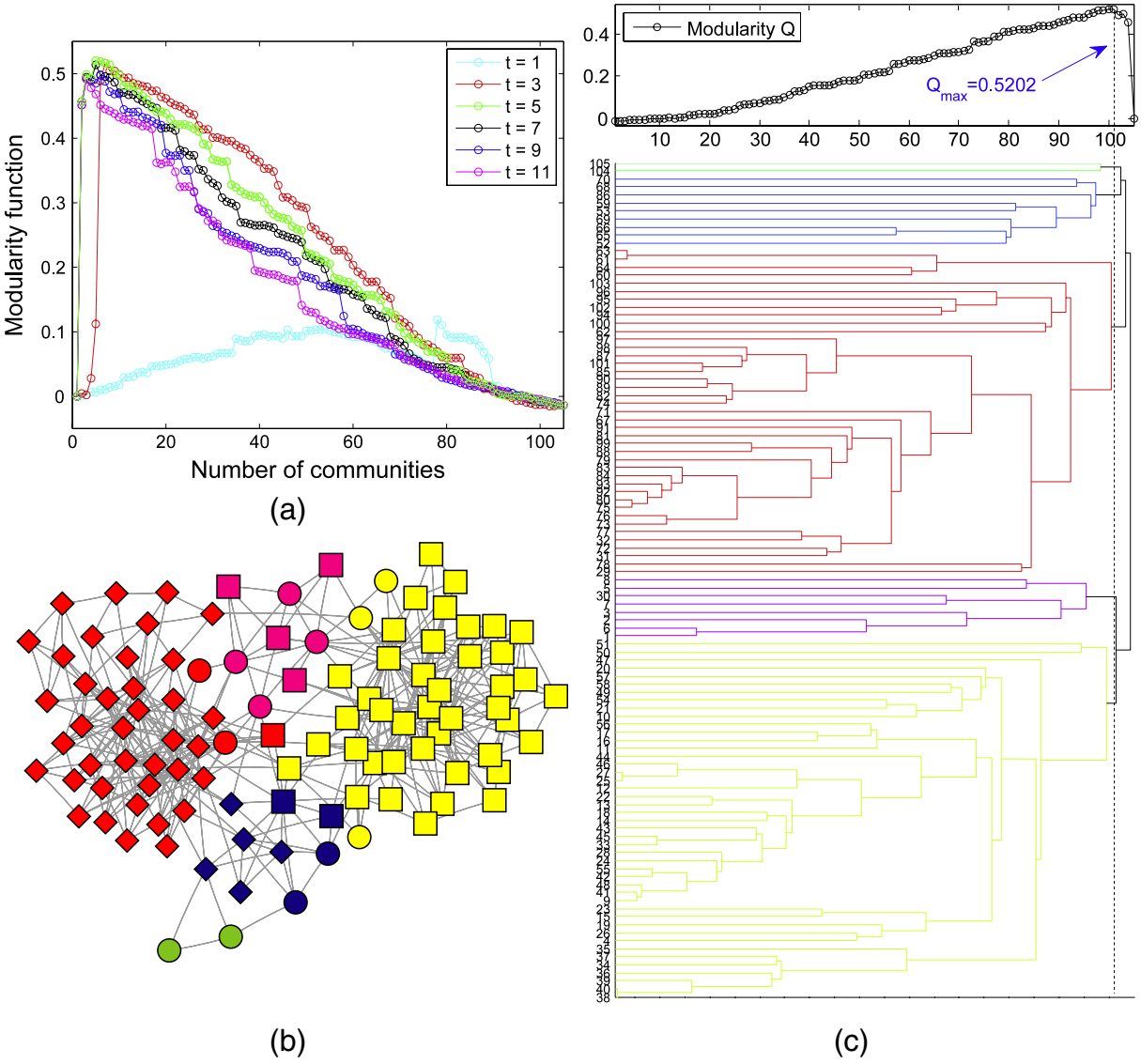

Figure 8. The computational results for the political books network detected by our method. (a) The modularity change with number of communities in each iteration for different time parameter t. (b) The community structure identified by setting $t = 5$ corresponds to the five communities represented by the colors. (c) The dendrogram of the hierarchical structures and the optimal partition with a maximal modularity $Q = 0.5202$ is denoted by a vertical dashed line.

nodes have been given whether they are conservative (box) or liberal (diamond), except for a small number of books that are neutral (ellipse). The modularity function Q change with the number of communities N in each agglomerative step for different time parameter t is shown in figure 8(a). The parameter is set by $t = 5$ in this model computation, and the community structure detected by our method, shown in figure 8(b), corresponds to five communities represented by the colors. In figure 8(c) we give the dendrogram of the hierarchical structures. The optimal partition with a maximal modularity $Q = 0.5202$

J. Stat. Mech. (2010) P12030

is denoted by a vertical dashed line. It seems that one of these communities consists almost entirely of liberal books and one almost entirely of conservative books. Most of the neutral books fall in the three remaining communities. Thus these books appear to form communities of co-purchasing that align closely with political views. The comparison with the linkage choices shown in table 1 and figure 6 indicates that our method gives good results.

4.2.4. The football team network. The last network we investigated is the college football network, which represents the game schedule of the 2000 season of Division I of the US college football league [10, 13], [19]–[22], [27, 29, 34, 36]. The nodes in the network represent 115 teams and edges represent regular season games between the two teams they connect. The teams are divided into conferences containing around 8 to 12 teams each. Games are more frequent between members of the same conference than between members of different conferences. The modularity function Q change with the number of communities N in each agglomerative step for different time parameter t is shown in figure 9(a). The parameter is set by $t = 1$ in this model computation, and the community structure detected by our method, shown in figure 9(b), corresponds to 11 communities represented by the colors. In figure 9(c) we give the dendrogram of the hierarchical structures. The optimal partition with a maximal modularity $Q = 0.6042$, which is much larger than the linkage choices shown in table 1, is denoted by a vertical dashed line. According to the results, we identify the community structure with a high degree of accuracy in that almost all of the football teams are correctly clustered with the others in their conference. The teams in the Independents conference seem not to belong to any community, but they tend to be clustered with the conference with which they are most closely associated. The Sunbelt conference is split into two communities, one is clustered with a team less connected in the Western Athletic conference and the other is clustered with the SEC conference. Only one member in Conference USA is grouped with most of the teams in the Western Athletic conference. All the other communities coincide with the known structure.

4.2.5. The SFI collaboration network. The last example is the collaboration network of scientists at the Santa Fe Institute, an interdisciplinary research center in Santa Fe, New Mexico [10, 19, 20]. The 271 nodes in this network represent scientists in residence at the Santa Fe Institute, during any part of calendar year 1999 or 2000, and their collaborators. A weighted edge is drawn between a pair of scientists if they coauthored one or more articles during the same time period. In figure 10(b), we illustrate the results from the application of our algorithm to the largest component of the collaboration graph, which consists of 118 scientists. The modularity function Q change with the number of communities N in each agglomerative step for different time parameter t is shown in figure 10(a). The parameter is set by $t = 9$ in this model computation, and the community structure detected by our method, shown in figure 10(b), corresponds to six communities represented by the colors. In figure 10(c) we give the dendrogram of the hierarchical structures. The optimal partition with a maximal modularity $Q = 0.7266$, which is much larger than the linkage choices shown in table 1 and figure 6, is denoted by a vertical dashed line. We find that our method splits the network into six communities with the divisions running principally along disciplinary lines. The community at the top of the figure

J. Stat. Mech. (2010) P12030

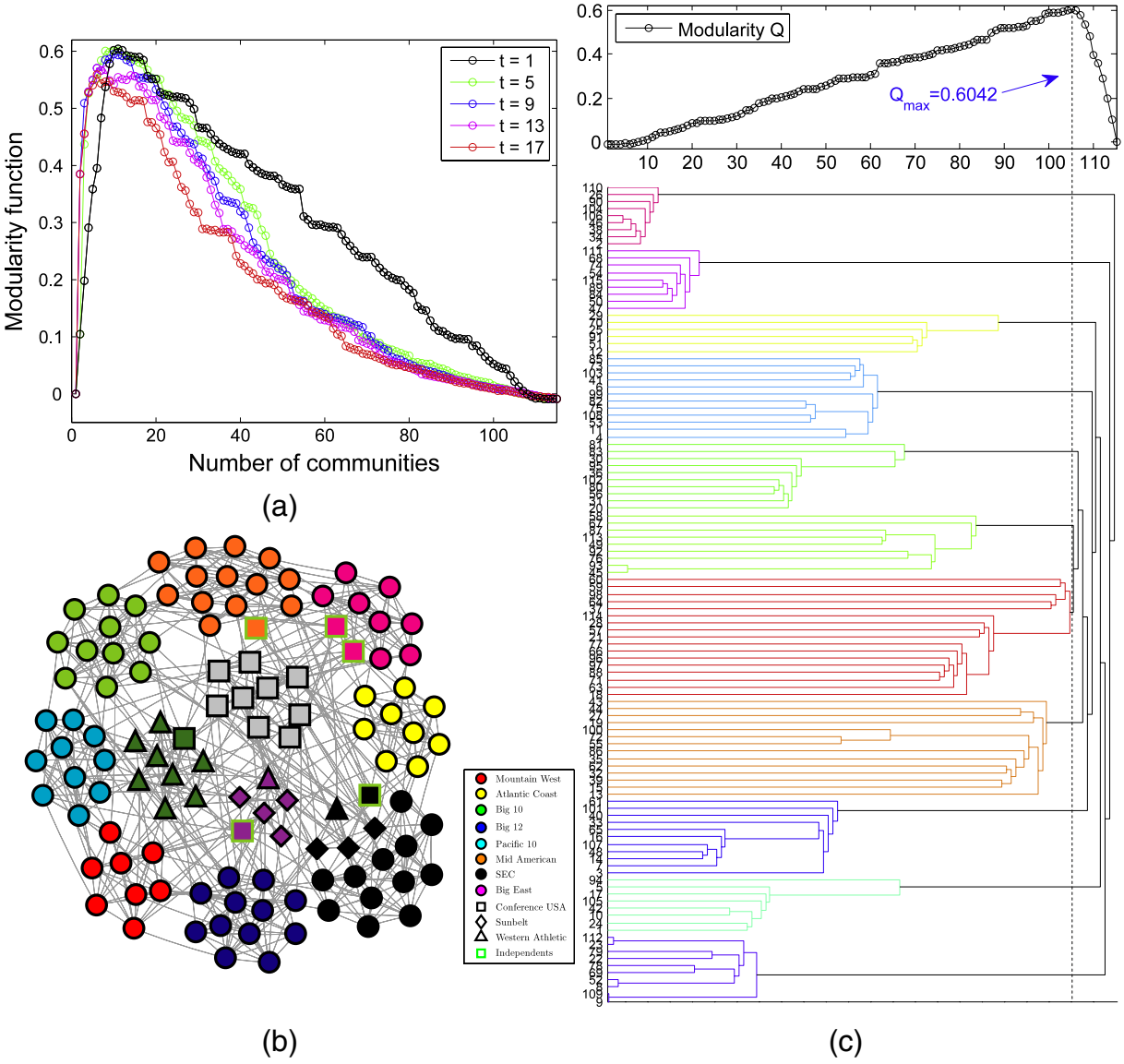

Figure 9. The computational results for the football team network detected by our method. (a) The modularity change with number of communities in each iteration for different time parameter t. (b) The community structure identified by setting $t = 1$ corresponds to 11 communities represented by the colors. (c) The dendrogram of the hierarchical structures and the optimal partition with a maximal modularity $Q = 0.6042$ is denoted by a vertical dashed line.

(yellow) represents a group of scientists using agent-based models to study problems in economics and traffic flow. The next community (magenta) represents a group of scientists working on mathematical models in ecology, and forms a fairly cohesive structure. The largest community (cyan, green, red) is a group working primarily in statistical physics, and seems sub-divided into three smaller groups. In this case, each sub-community seems to revolve around the research interests of one dominant member. The final community at the bottom of the figure (blue) is a group working primarily on the structure of RNA.

J. Stat. Mech. (2010) P12030

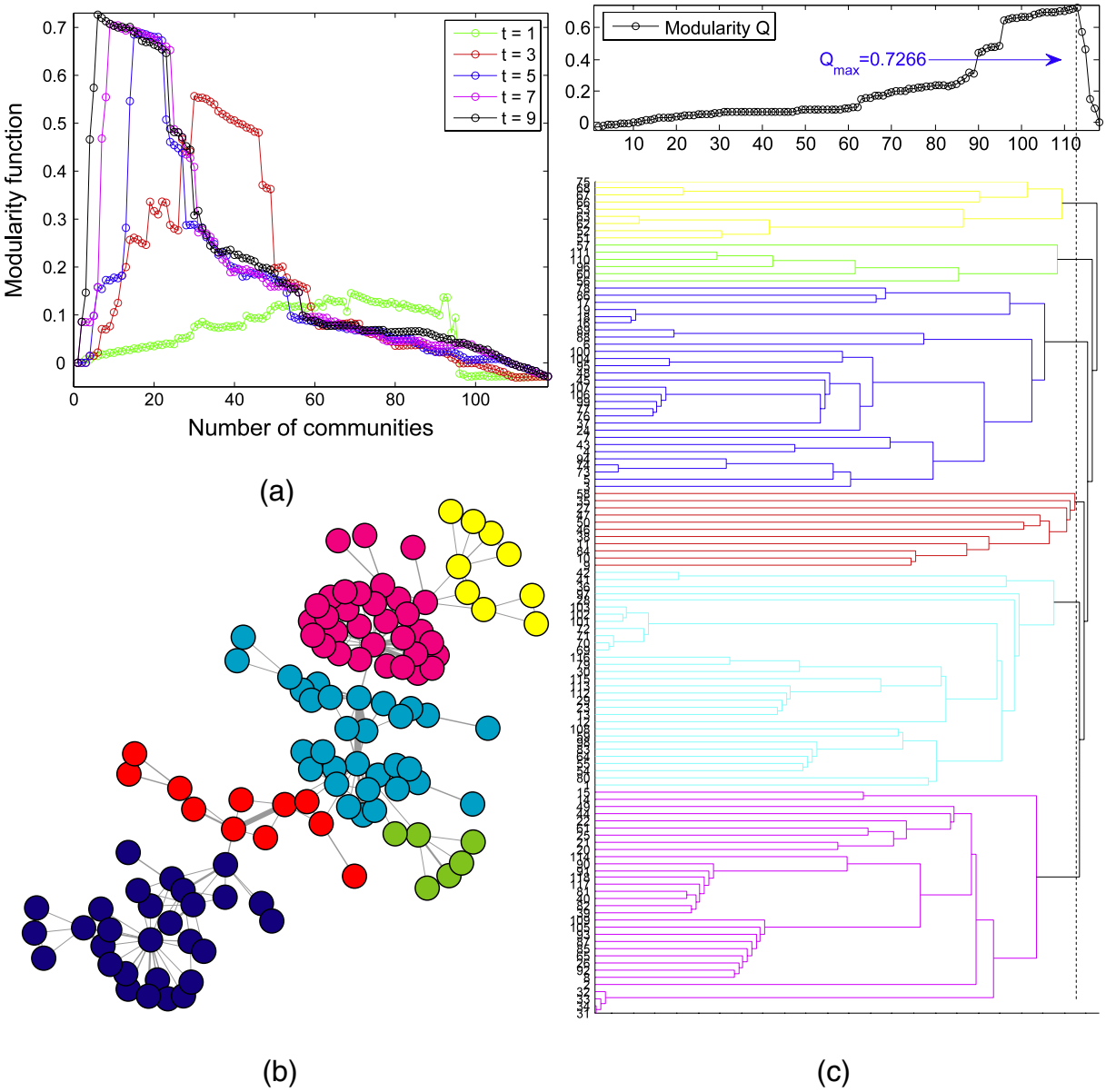

Figure 10. The computational results for the SFI collaboration network detected by our method. (a) The modularity change with number of communities in each iteration for different time parameter t. (b) The community structure identified by setting $t = 9$, corresponds to six communities represented by the colors. (c) The dendrogram of the hierarchical structures and the optimal partition with a maximal modularity $Q = 0.7266$ is denoted by a vertical dashed line.

5. Conclusions

In this paper, we extend the measure of diffusion distance between nodes in a complex network to a generalized form on the coarse-grained network, whose nodes are the communities of the original network, with data parameterization via eigenmaps. This notion of proximity of 'nodes' in the coarse-grained networks reflects the intrinsic geometry

J. Stat. Mech. (2010) P12030

of the meta-node set in terms of connectivity of the communities in a diffusion process. This metric is usually more appropriate than the linkage choices in traditional clustering literature, as it takes into account all the evidence relating the two communities. Furthermore, since diffusion-based distances add up the contribution from all the possible paths, they are also robust to noise. Nodes are then grouped into communities through an agglomerative hierarchical clustering technique and the modularity function is used to select the best partition of the resulting dendrogram. The simulated experiments on artificial networks show very satisfactory results in that the agglomerative process can efficiently identify the community structure of a given network and the number of communities can be automatically determined without any prior knowledge about the community structure. Moreover, successful application to several real-world networks confirm the effectiveness of the present algorithm.

Acknowledgments

We thank Professor S Fortunato for kindly sharing the codes for generating the LFR benchmarks. We are also grateful to Professor M E J Newman and Professor H Zhou for providing the data for the karate club network, the dolphins network, the American political books network, the football team network and the SFI collaboration network. This work is supported by the Natural Science Foundation of China under Grant 10871010 and the National Basic Research Program of China under Grant 2005CB321704.

References

[1] Albert R and Barabási A-L, *Statistical mechanics of complex networks*, 2002 *Rev. Mod. Phys.* **74** 47
[2] Newman M E J, *The structure and function of complex networks*, 2003 *SIAM Rev.* **45** 167
[3] Newman M E J, Barabási A-L and Watts D J, 2005 *The Structure and Dynamics of Networks* (Princeton, NJ: Princeton University Press)
[4] Barabási A L, Jeong H, Neda Z, Ravasz E, Schubert A and Vicsek T, *Evolution of the social network of scientific collaborations*, 2002 *Physica* A **311** 590
[5] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabási A-L, *Hierarchical organization of modularity in metabolic networks*, 2002 *Science* **297** 1551
[6] Flake G W, Lawrence S, Giles C L and Coetzee F M, *Self-organization and identification of web communities*, 2002 *IEEE Comput.* **35** 66
[7] Shi J and Malik J, *Normalized cuts and image segmentation*, 2000 *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888
[8] Meilă M and Shi J, *A random walks view of spectral segmentation*, 2001 *Proc. 8th Int. Workshop on Artificial Intelligence and Statistics (Kaufmann, San Francisco)* p 92
[9] Lafon S and Lee A B, *Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization*, 2006 *IEEE Trans. Pattern. Anal. Mach. Intell.* **28** 1393
[10] Girvan M and Newman M E J, *Community structure in social and biological networks*, 2002 *Proc. Nat. Acad. Sci.* **99** 7821
[11] Newman M E J and Girvan M, *Finding and evaluating community structure in networks*, 2004 *Phys. Rev.* E **69** 026113
[12] Clauset A, Newman M E J and Moore C, *Finding community structure in very large networks*, 2004 *Phys. Rev.* E **70** 066111
[13] Newman M E J, *Fast algorithm for detecting community structure in networks*, 2004 *Phys. Rev.* E **69** 066133
[14] Newman M E J, *Detecting community structure in networks*, 2004 *Eur. Phys. J.* B **38** 321
[15] Newman M E J, *Finding community structure in networks using the eigenvectors of matrices*, 2006 *Phys. Rev.* E **74** 036104
[16] Newman M E J, *Modularity and community structure in networks*, 2006 *Proc. Nat. Acad. Sci.* **103** 8577

J. Stat. Mech. (2010) P12030

[17] Newman M E J and Leicht E A, *Mixture models and exploratory analysis in networks*, 2007 *Proc. Nat. Acad. Sci.* **104** 9564
[18] Noh J D and Rieger H, *Random walks on complex networks*, 2004 *Phys. Rev. Lett.* **92** 118701
[19] Zhou H, *Network landscape from a Brownian particle's perspective*, 2003 *Phys. Rev.* E **67** 041908
[20] Zhou H, *Distance, dissimilarity index, and network community structure*, 2003 *Phys. Rev.* E **67** 061901
[21] Pons P and Latapy M, *Computing communities in large networks using random walks*, 2006 *J. Graph Algorithms Appl.* **10** 191–218
[22] Hu Y, Li M, Zhang P, Fan Y and Di Z, *Community detection by signaling on complex networks*, 2008 *Phys. Rev.* E **78** 016115
[23] Rosvall M and Bergstrom C T, *Maps of random walks on complex networks reveal community structure*, 2008 *Proc. Nat. Acad. Sci.* **105** 1118
[24] Weinan E, Li T and Vanden-Eijnden E, *Optimal partition and effective dynamics of complex networks*, 2008 *Proc. Nat. Acad. Sci.* **105** 7907
[25] Li T, Liu J and Weinan E, *Probabilistic framework for network partition*, 2009 *Phys. Rev.* E **80** 026106
[26] Wu F and Huberman B A, *Finding communities in linear time: a physics approach*, 2004 *Eur. Phys. J.* B **38** 331
[27] Reichardt J and Bornholdt S, *Detecting fuzzy community structures in complex networks with a Potts model*, 2004 *Phys. Rev. Lett.* **93** 218701
[28] Duch J and Arenas A, *Community detection in complex networks using extremal optimization*, 2005 *Phys. Rev.* E **72** 027104
[29] Hofman J M and Wiggins C H, *Bayesian approach to network modularity*, 2008 *Phys. Rev. Lett.* **100** 258701
[30] Arenas A, Fernandez A and Gomez S, *Analysis of the structure of complex networks at different resolution levels*, 2008 *New J. Phys.* **10** 053039
[31] Blondel V D, Guillaume J L, Lambiotte R and Lefebvre E, *Fast unfolding of communities in large networks*, 2008 *J. Stat. Mech.* P10008
[32] Shen H W, Cheng X Q and Guo J F, *Quantifying and identifying the overlapping community structure in networks*, 2009 *J. Stat. Mech.* P07042
[33] Cheng X Q and Shen H W, *Uncovering the community structure associated with the diffusion dynamics on networks*, 2010 *J. Stat. Mech.* P04024
[34] Zhang S, Wang R S and Zhang X S, *Identification of overlapping community structure in complex networks using fuzzy c-means clustering*, 2007 *Physica* A **374** 483
[35] Shen H, Cheng X, Cai K and Hu M B, *Detect overlapping and hierarchical community structure in networks*, 2009 *Physica* A **388** 1706
[36] Liu J and Liu T, *Detecting community structure in complex networks using simulated annealing with k-means algorithms*, 2010 *Physica* A **389** 2300
[37] Danon L, Diaz-Guilera A, Duch J and Arenas A, *Comparing community structure identification*, 2005 *J. Stat. Mech.* P09008
[38] Lancichinetti A, Fortunato S and Radicchi F, *Benchmark graphs for testing community detection algorithms*, 2008 *Phys. Rev.* E **78** 046110
[39] Lancichinetti A and Fortunato S, *Community detection algorithms: a comparative analysis*, 2009 *Phys. Rev.* E **80** 056117
[40] Fortunato S, *Community detection in graphs*, 2010 *Phys. Rep.* **486** 75
[41] Schilders W H A, Van der Vorst H A and Rommes J, 2008 *Model Order Reduction: Theory, Research Aspects and Applications* (Berlin: Springer)
[42] Hastie T, Tibshirani R and Friedman J, 2001 *The Elements of statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer)
[43] Lovasz L, *Random walks on graphs: a survey*, 1993 *Combin. Paul Erdös Eighty* **2** 1
[44] Chung F R K, 1997 *Spectral Graph Theory* (Rhode Island: American Mathematical Society)
[45] Fortunato S and Barthélemy M, *Resolution limit in community detection*, 2007 *Proc. Nat. Acad. Sci.* **104** 36
[46] Good B H, De Montjoye Y A and Clauset A, *Performance of modularity maximization in practical contexts*, 2010 *Phys. Rev.* E **81** 046106
[47] Zachary W W, *An information flow model for conflict and fission in small groups*, 1977 *J. Anthrop. Res.* **33** 452
[48] Lusseau D, *The emergent properties of a dolphin social network*, 2003 *Proc. R. Soc.* B **270** 186
[49] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M, *The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations*, 2003 *Behav. Ecol. Sociobiol.* **54** 396

J. Stat. Mech. (2010) P12030