Quantitative function for community detection

Zhenping Li,[1,2,*] Shihua Zhang,[2,3,*] Rui-Sheng Wang,[4] Xiang-Sun Zhang,[2,†] and Luonan Chen[5,6,†]
[1]*Beijing Wuzi University, Beijing 101149, China*
[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China*
[3]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China*
[4]*School of Information, Renmin University of China, Beijing 100872, China*
[5]*Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[6]*Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka 574-8530, Japan*
(Received 6 October 2007; revised manuscript received 2 December 2007; published 10 March 2008)

We propose a quantitative function for community partition—i.e., modularity density or *D* value. We demonstrate that this quantitative function is superior to the widely used modularity *Q* and also prove its equivalence with the objective function of the kernel *k* means. Both theoretical and numerical results show that optimizing the new criterion not only can resolve detailed modules that existing approaches cannot achieve, but also can correctly identify the number of communities.

DOI: 10.1103/PhysRevE.77.036109 PACS number(s): 89.75.Hc, 87.23.Ge

I. INTRODUCTION

It has been widely demonstrated in the past that many interesting systems can be represented as networks composed of vertices and edges [1–3]. Such systems include the internet, social and friendship networks, food webs, biomolecular networks, and citation networks. The prolific progress in the study of complex networks driven by the development of information technology and the increasing availability of huge networked data in the real world have revealed many interesting topological properties, such as small-world properties, power-law degree distributions, and network motifs.

A topic of great interest in the area of complex networks is the community structure and its detection. A community could be roughly described as a collection of vertices in a subgraph that are densely connected among themselves while being loosely connected to the vertices outside the subgraph. Since many networks exhibit such a community structure, the characterization and detection of such a community structure have great practical significance. Taking biological molecular networks as an example, dividing protein interaction networks into modular groups provides strong evidence of independent functions and actions for proteins in different subgraphs [3,4].

There have been abundant techniques proposed to detect community structure [5,6] and fuzzy community structure [7–10] in a network from various fields, but most methods require a definition of community that imposes the limit up to which a group should be considered as a community. However, the concept of community itself is qualitative; e.g., nodes must be more connected within its community than with the rest of the network. Therefore, its quantification is still a subject of debate. Two aspects greatly complicate this problem. In general, the size heterogeneity of communities often greatly affects the measure of community [11]. Another aspect is that, even in a specific network, the generation mechanism or link degree may vary greatly.

A widely used quantitative measure for evaluating the partition of a network is called modularity (known as *Q*), which was introduced by Newman and Girvan [12]. If one chooses the modularity as the relevant quantitative function, the problem of community detection becomes equivalent to modularity optimization. Modularity optimization seems to be an effective method to detect communities both in real and in artificially generated networks. By defining *Q* as an objective function, a class of methods aiming to maximize the modularity has been developed [13–15]. However, the modularity has been exposed to resolution limits [16–18]. Fortunato and Barthélemy [16] recently claimed that modularity contains an intrinsic scale that depends on the total size of links in the network. Modules smaller than this scale may not be resolved even in the extreme case that they are complete graphs connected by single bridges [16]. Similar observations have also been raised by [17,18]. In [18], a generalized modularity called localized modularity measure was proposed.

In this paper, we propose a quantitative measure for evaluating the partition of a network into communities based on the concept of average modularity degree. We call this quantitative measure the modularity density or *D* value. In addition to the simple form, we show that the proposed criterion improves the resolution limit in community detection based on theoretical analysis and numerical test of artificial networks and real-world networks. We also theoretically reveal the equivalence of modularity density and the objective function of kernel *k* means, which explains the implication of the criterion in another way.

II. MODULARITY DENSITY

Given a network $G=(V,E)$, V is the vertex set, E is the edge set, and A is the adjacent matrix of G. If V_1 and V_2 are two disjoint subsets of V, we further define $L(V_1,V_2)=\Sigma_{i\in V_1,j\in V_2}A_{ij}$, $L(V_1,V_1)=\Sigma_{i\in V_1,j\in V_1}A_{ij}$, and $L(V_1,\bar{V}_1)=\Sigma_{i\in V_1,j\in \bar{V}_1}A_{ij}$, where $\bar{V}_1=V-V_1$. Given the partition of a network G, $G_1(V_1,E_1),\ldots,G_m(V_m,E_m)$, where V_i and E_i are,

*The first two authors contributed equally to this paper.
†Corresponding authors.

respectively, the node set and the edge set of G_i for $i=1,\ldots,m$, the well-known modularity Q is defined as follows:

$$Q=\sum_{i=1}^{m}\left[\frac{L(V_i,V_i)}{L(V,V)}-\left(\frac{L(V_i,V)}{L(V,V)}\right)^2\right]. \quad (1)$$

Modularity optimization for Q seems to be an effective method to detect communities in networks. However, Fortunato and Barthélemy [16] recently pointed out the serious resolution limits of this method and claimed that the size of a detected module depends on the size of the whole network. This is mainly because the modularity measure does not contain information on the number of nodes in a community and the choice of partition is highly sensitive to the total number of links in the network [17]. In the following, we will introduce a measure D, which is related to the density of subgraphs to overcome this problem. We first define the average modularity degree of subgraph $G_i(V_i,E_i)$ as follows:

$$d(G_i)=d_{in}(G_i)-d_{out}(G_i),$$

where $d_{in}(G_i)$ is the average inner degree of the subgraph G_i, which is equal to twice the number of edges in subgraph G_i divided by the number of nodes in set V_i. $d_{out}(G_i)$ is the average outer degree of subgraph G_i, which is equal to the number of edges with one node in V_i and the other node outside V_i divided by the number of nodes in V_i. It can be easily formulated as

$$d(G_i)=\frac{L(V_i,V_i)-L(V_i,\bar{V}_i)}{|V_i|}.$$

The intuitive idea is that $d(G_i)$ should be as large as possible for a valid "community." Then we define the modularity density of a partition as the sum of all average modularity degrees of G_i for $i=1,\ldots,m$. Let D denote the modularity density (called the D value in this paper) of a partition of a network G into communities $G_1,\ldots,G_m$. Then, in contrast to Q, D can be calculated as follows:

$$D=\sum_{i=1}^{m}d(G_i)=\sum_{i=1}^{m}\frac{L(V_i,V_i)-L(V_i,\bar{V}_i)}{|V_i|}. \quad (2)$$

The summation extends to all communities G_i of a given partition. Note that this measure provides a way to determine if a certain mesoscopic description of the graph is accurate in terms of communities. The larger the value of D, the more accurate a partition is. So the community-detection problem can be viewed as a problem of finding a partition of a network such that its modularity density D is maximized. Since our purpose is to maximize the modularity density D, every term $d(G_i)$ must be non-negative. Therefore, the partition (subgraphs) by optimizing D results in communities consistent with the weak definition suggested by Radicchi *et al.* [19].

The search for optimal modularity density D is a NP-hard problem due to the fact that the space of possible partitions grows faster than any power of system size. In this paper we will prove that the modularity-detection problem based on optimizing D value is equal to a kernel k-means problem. Such a theoretical result may be exploited to derive an efficient computational algorithm for optimizing D.

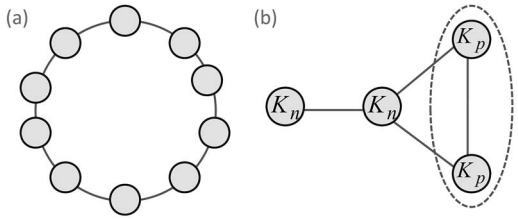

FIG. 1. Schematic examples. (a) The clique circle graph in the left figure. Each module is a clique of n nodes, and two adjacent modules are connected by one edge. (b) A network with two pairs of identical cliques in the right figure. One pair of cliques have n nodes, and the other pair of cliques have p nodes.

III. IMPROVING RESOLUTION LIMITS BY MODULARITY DENSITY

Although detecting communities based on the optimization of modularity has been widely used as a popular method, Fortunato and Barthélemy recently found that modularity optimization may fail to identify modules smaller than a scale even in cases where modules are unambiguously defined [16]. This scale depends on the total size of the network and on the degree of interconnectedness of the modules. In this paper, we propose a modularity density D to overcome such a problem. To assess the reliability of modularity density, we perform the same tests as those examples from Fortunato and Barthélemy [16].

A. Modularity density does not divide a clique into two parts

Given a clique with n nodes, maximizing modularity density or D does not divide it into two or more parts.

We can prove this result by contradiction. Suppose that P is a partition which divides the clique into G_1 and G_2 and the number of nodes in G_1 and G_2 are n_1 and n_2, respectively; then, the number of edges between G_1 and G_2 is n_1n_2. Let D_0 be the modularity density of G and let D_1 denote the modularity density of partition P; then,

$$D_0=n-1,$$

$$D_1=\frac{n_1(n_1-1)-n_1n_2}{n_1}+\frac{n_2(n_2-1)-n_1n_2}{n_2}=-2.$$

Since $D_0>D_1$, maximizing D value does not divide the clique into two parts.

B. Modular density can resolve most modular networks correctly

To test the quality of the modularity density, we use the schematic example from [16], which is a network consisting of a ring of cliques connected through single links [see Fig. 1(a)]. Each clique is a complete graph K_n with n $(n\geq 3)$ nodes and $n(n-1)/2$ links. Assuming that there are m cliques

($m \geq 2$ can be exactly divided by k, where $k \geq 2$ is an integer), the network has a total of $N=mn$ nodes and $L=mn(n-1)/2+m$ links.

The network has a clear modular structure where each community corresponds to a single clique, but the correct result cannot be obtained by optimizing Q value [16]. Now we optimize D value to find the solution. The modularity density D_{single} of the natural partition can be easily and analytically calculated as follows:

$$D_{single} = m\frac{n(n-1)-2}{n} = m\left(n-1-\frac{2}{n}\right).$$

On the other hand, the modularity density D_k of the partition in which the k consecutive cliques are considered as single communities is

$$D_k = \frac{m}{k}\frac{kn(n-1)+2(k-3)}{kn}.$$

Supposing $k \geq 2$, $n \geq 3$, and $m \geq 2$; then,

$$\begin{aligned} D_{single} - D_k &= m\left(n-1-\frac{2}{n}\right) - \frac{m}{k}\frac{kn(n-1)+2(k-3)}{kn} \\ &= m\left[(n-1)-\frac{2}{n}-\frac{n-1}{k}-\frac{2(k-3)}{k^2 n}\right] \\ &> m\left[(n-1)-\frac{2}{n}-\frac{n-1}{k}-\frac{2}{kn}\right] \\ &\geq m\left[(n-1)-\frac{2}{n}-\frac{n-1}{2}-\frac{1}{n}\right] \\ &= m\left[(n-1)-\frac{3}{n}-\frac{n-1}{2}\right] > 0. \end{aligned}$$

Although the above analysis is conducted for the special partition that the k consecutive cliques are considered as single communities, by a similar argument we can prove that such a result is actually valid for any kind of grouping cliques (i.e., any combination of cliques as communities). Therefore, these results, along with the fact that optimizing D does not divide a clique into two parts, lead to the conclusion that the maximal value of D exactly corresponds to the correct partition (with each single clique as a community). In other words, optimizing D can lead to the correct partition. A complete analytical proof based on optimization method leads to the same conclusion (see Appendix B).

For the special case $n=2$, the network is a circle of $2m$ nodes and $2m$ links. Suppose that m can be exactly divided by k; then, the modularity density D_k of the partition that the $2k$ consecutive nodes constitute an individual community is

$$D_k = \frac{m}{k}\frac{4(k-1)}{2k} = 2m\frac{k-1}{k^2}.$$

It is easy to verify that the maximal value of D_k is obtained when $k=2$; i.e., every partition is a path with four nodes. The reason that this result does not agree with above one is that K_2 is a trivial clique. Since every K_2 is a single edge with the same inner degree as the outer degree, it cannot be a community by itself.

C. Modular density can detect communities with different sizes

Suppose that there is a network consisting of four cliques, two of which are K_n and the other two are K_p, for $3 \leq p \leq n$ [see Fig. 1(b)].

In [16,18], the authors observed that optimizing Q has a tendency to merge small modules. In the following, we prove that the D value based optimization does not have such a problem.

Let $D_{separate}$ denote the modularity density of the partition in which the two small cliques are separated and D_{merge} denote the modularity density of the partition where the two small cliques are merged; then,

$$\begin{aligned} D_{separate} &= \frac{n(n-1)-1}{n} + \frac{n(n-1)-3}{n} + 2\frac{p(p-1)-2}{p} \\ &= \frac{n(n-1)-1}{n} + \frac{n(n-1)-3}{n} + 2(p-1) - \frac{4}{p}, \end{aligned}$$

$$\begin{aligned} D_{merge} &= \frac{n(n-1)-1}{n} + \frac{n(n-1)-3}{n} + \frac{2p(p-1)}{2p} \\ &= \frac{n(n-1)-1}{n} + \frac{n(n-1)-3}{n} + (p-1). \end{aligned}$$

It is easy to verify that when $p \geq 3$,

$$D_{separate} - D_{merge} = 2(p-1) - \frac{4}{p} - (p-1) > 0.$$

The above analysis is conducted for the special partition that two small cliques are merged as a community with each other clique as a community. With the fact that optimizing D does not partition a clique into two parts, it is easy to see that any other partition has a lower D value than the one with each clique as a community. Therefore, the optimal value of D corresponds to the correct partition. In contrast to the modularity Q, optimizing the D value can correctly detect communities with any sizes.

Based on the above discussion, clearly the maximum D value is often achieved when the network is correctly partitioned. Such a fact demonstrates the effectiveness of the D value acting as a quantitative function for community structure.

IV. EQUIVALENCE OF MODULARITY DENSITY AND KERNEL k MEANS

Once the number of communities is fixed, the optimization process of modularity density leads to the detection of proper communities and the quality of the solution is evaluated on the basis of its D value. On the other hand, the efficiency of optimizing modularity density can be exploited based on the equivalence of modularity density and the objective function of kernel k means [20]. Next, we derive such theoretical results.

Given a set of data vectors $V=\{v_i\}_{i=1}^N$ with $v_i \in \mathbb{R}^n$, the goal of kernel k means is to find an m-way disjoint partition $\{V_c\}_{c=1}^m$ of the data (where V_c represents the cth cluster) such that the following objective function is minimized:

$$F = \sum_{c=1}^{m} \sum_{v_i \in V_c} \|\phi(v_i) - m_c\|^2, \tag{3}$$

where $F=F(\{V_c\}_{c=1}^m)$ and

$$m_c = \frac{\sum_{v_i \in V_c} \phi(v_i)}{|V_c|},$$

$|V_c|$ is the cardinality of the subset V_c, and ϕ is a function mapping the vectors in V onto a generally higher-dimensional space. Clearly, if ϕ is the identity function, the above equation recovers the standard definition of the k means.

We can easily obtain the following formulation by expanding the distance term $\|\phi(v_i)-m_c\|^2$ in the objective function:

$$\|\phi(v_i) - m_c\|^2 = \phi(v_i) \cdot \phi(v_i) - \frac{2 \sum_{v_j \in V_c} \phi(v_i) \cdot \phi(v_j)}{|V_c|} + \frac{\sum_{v_j \in V_c} \sum_{v_l \in V_c} \phi(v_j) \cdot \phi(v_l)}{|V_c|^2}. \tag{4}$$

Notice that only the inner products are used in the equation. As a result, for a given kernel matrix K, where $K_{ij} = \phi(v_i) \cdot \phi(v_j)$, we can compute the distances between two data points v_i and v_j without knowing explicit representations of $\phi(v_i)$ and $\phi(v_j)$. It has been shown that any positive semidefinite matrix K can be thought of as a kernel matrix [21].

Using the kernel matrix, Eq. (3) can be rewritten as

$$F = \sum_{c=1}^{m} \sum_{v_i \in V_c} \left(K_{ii} - \frac{2 \sum_{v_j \in V_c} K_{ij}}{|V_c|} + \frac{\sum_{v_j \in V_c} \sum_{v_l \in V_c} K_{jl}}{|V_c|^2} \right). \tag{5}$$

On the other hand, the purpose of this paper is to look for the m-way disjoint partition $\{V_c\}_{c=1}^m$ of V that maximizes the modularity density:

$$D = \sum_{c=1}^{m} \frac{L(V_c, V_c) - L(V_c, V_c)}{|V_c|}, \tag{6}$$

where $D=D(\{V_c\}_{c=1}^m)$. Let us first define a diagonal degree matrix C with $C_{ii}=\Sigma_{j=1}^n A_{ij}$. Then we associate the given graph with an $N \times N$ kernel matrix as follows:

$$K = \sigma I + 2A - C, \tag{7}$$

where I is the identity matrix and σ is a real number chosen to be sufficiently large so that the K is positively definite. Now given an m-way disjoint partition $\{V_c\}_{c=1}^m$ of the graph, the corresponding modularity density and the objective function of kernel k means are related as follows:

$$F = (N-m)\sigma - D. \tag{8}$$

An important point follows: F attains its minimum if and only if the maximum of D is achieved, independently of σ, as is shown in [20], when considering the standard iterations of k means. Therefore, kernel k means may be straightforwardly used to find the m optimal clusters of the graph by simply maximizing the modularity density. On the other hand, similar to Q, we can also use k means to determine an appropriate m, e.g., by varying m so as to obtain an optimal objective function.

Furthermore, if we use the following kernel matrix K_λ instead of K in Eq. (7),

$$K_\lambda = \sigma I + 2[\lambda A - (1-\lambda)(C-A)], \quad 0 \le \lambda \le 1, \tag{9}$$

we can obtain a more general modularity density measure:

$$D_\lambda = \sum_{i=1}^{m} \frac{2\lambda L(V_i, V_i) - 2(1-\lambda) L(V_i, \bar{V}_i)}{|V_i|}. \tag{10}$$

When $\lambda=1$, D_λ is equivalent to the ratio association [20]; when $\lambda=0$, D_λ is equivalent to the ratio cut [20]; when $\lambda=0.5$, D_λ is equivalent to the modularity density D. So the general modularity density D_λ can be viewed as a combination of the ratio association and the ratio cut. Generally, optimization of the ratio association algorithm often divides a network into small communities [22], while optimization of the ratio cut often divides a network into large communities. The general modularity density D_λ, which is a convex combination of these two indexes, can avoid the resolution limits. In other words, we can decompose the network into large communities and small communities by using a small λ and a large λ, respectively. As a matter of fact, the phenomenon of multiple resolutions for modular structures in complex networks is natural [23]. Many complex networks have a hierarchical or nesting community structure [24]. Therefore, generally there is no absolutely "optimal" standard for the community structure of complex networks, which means that we cannot obtain the so-called optimal λ value in a general sense. In other words, this general function can be applied to analyze the topological structure and uncover more detailed and hierarchical organization of complex systems by varying the λ value. However, for specific cases, we may obtain the optimal or "appropriate" D_λ to find out community structure by exploiting additional information on the topological structure of networks as well as the context implication of communities in networks

V. EXPERIMENTAL RESULTS

In this section, we conduct experiments on both artificial networks and well-studied real networks. We first formulate the community-detection problem into an integer programming model to optimize D value (see Appendix A), and then the integer programming is solved by the LINGO software.

A. Artificial networks

First, we do the test on the computer-generated networks. Each network has 128 nodes, which are divided into 4 com-

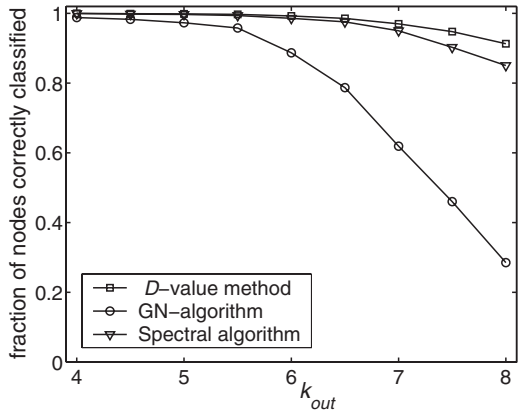

FIG. 2. Test of various methods on computer-generated networks with known community structures. It is a plot of the fraction of nodes correctly classified with respect to k_{out}. Each point is an average over 100 realizations of the networks.

munities with 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 edge connections k_{out} of each node to nodes of other modules. The experiment was designed by Girvan and Newman [25] and has been broadly used to test community-detection algorithms [25,26].

The computational results for this experiment are summarized in Fig. 2 and Table I, which show the fraction of nodes that are correctly classified into the communities with respect to k_{out} by our method and the other algorithms, respectively. We can see that our method based on the D value performs much better than other algorithms, such as the Girvan-Newman (GN) algorithm [25] and the spectral algorithm based on optimizing Q [14]. Table I demonstrates the results of the cluster compression algorithm, Q optimization algorithm, and D optimization algorithm. From Table I we can see that, when the communities are of equal size and similar total degree, every method performs very well. At the same time, when $k_{out}=8$, which indicates that the corresponding networks are difficult to be partitioned, our method has the highest accuracy. When the communities vary in size or in total degree, the modularity optimization approach is more difficult to resolve the community structure (Table I) [17].

TABLE I. Benchmark performance for symmetric and asymmetric group detection measured as fraction of correct assignments, averaged over 100 network realizations with the standard deviation in parentheses.

Group	k_{out}	Compression	Q	D value
Symm.	6	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	7	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)
	8	0.87 (0.08)	0.89 (0.05)	0.91 (0.03)
Node asymm.	6	0.99 (0.01)	0.85 (0.04)	0.99 (0.01)
	7	0.96 (0.04)	0.80 (0.03)	0.98 (0.02)
	8	0.82 (0.10)	0.74 (0.05)	0.94 (0.03)
Link asymm.	2	1.00 (0.00)	1.00 (0.01)	1.00 (0.00)
	3	1.00 (0.00)	0.96 (0.03)	1.00 (0.00)
	4	1.00 (0.01)	0.74 (0.10)	0.99 (0.01)

We adopted the same method in [17] to construct asymmetric networks; i.e., three of the four groups in the benchmark test were merged to form a series of test networks, each with one large group of 96 nodes and one small group of 32 nodes. These asymmetrically sized networks are harder for both the Q optimization algorithm and cluster compression algorithm, but the D optimization algorithm can recover the underlying community structure more often than other two methods by a sizable margin. Finally, we conducted another set of benchmark tests using the link asymmetric networks used in [17]. They are composed of two groups, each with 64 nodes, but with different average degrees of 8 and 24 links per node. For these networks, we use $k_{out}=2,3,4$, for which the D optimization algorithm has a comparable result with the cluster compression algorithm and can recover community structure more often than the modularity optimization approach.

In general, before resolving the community structure, we must determine the number of communities in the network; then, we can partition the network into communities. This problem can be solved by using the extended modularity density D_λ. From the result of extensive simulation, we found that the maximum D value can often be obtained when the network is correctly partitioned. So we can determine the number of communities according to the D value; that is, the maximum D value corresponds to the correct number of communities. On the other hand, since the number of communities varies in different networks, we can use the extended D_λ instead of D to determine the number of communities. In this case, we can adjust parameter λ to obtain the proper number of communities. For example, we can use large λ to obtain communities of small size or use small λ to obtain communities of large size.

To test the performance of our method in selecting the number of communities, we do some simulations on the networks of Table I. Using proper λ, we can find the number of communities when D_λ is maximized. Then we summarize the results of our method and the results of the other two methods in Table II. From Table II, in any case, our method performs much better than the other two methods.

B. Real-world networks

1. Karate club network

Now we do the test on real networks. The first example is the famous karate club network analyzed by Zachary [27]. It 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 small ones. The question is that whether 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.

By using our method, the network was partitioned into two communities exactly consistent with real partition when $k=2$ (see Fig. 3). However, maximizing the D value, we

TABLE II. Benchmark performance for model selection measured as fraction of correct identification of number of groups, averaged over 100 network realizations with the average number of assigned modules in parentheses.

Group	k_{out}	Compression	Q	D value
Symm.	6	1.00 (4.00)	1.00 (4.00)	1.00 (4.00) ($\lambda=0.65$)
	7	1.00 (4.00)	1.00 (4.00)	1.00 (4.00) ($\lambda=0.65$)
	8	0.14 (1.93)	0.70 (4.33)	0.82 (4.18) ($\lambda=0.80$)
Node	6	1.00 (2.00)	0.00 (4.95)	1.00 (2.00) ($\lambda=0.65$)
asymm.	7	0.80 (1.80)	0.00 (4.97)	1.00 (2.00) ($\lambda=0.65$)
	8	0.06 (1.06)	0.00 (5.29)	0.68 (1.70) ($\lambda=0.65$)
Link	2	1.00 (2.00)	0.00 (3.10)	1.00 (2.00) ($\lambda=0.50$)
asymm.	3	1.00 (2.00)	0.00 (4.48)	1.00 (2.00) ($\lambda=0.50$)
	4	1.00 (2.00)	0.00 (5.55)	1.00 (2.00) ($\lambda=0.60$)

obtained the "optimal" partition with $k=4$ which is also reasonable from the topology of the network.

2. Football team network

The second real network is the college football network of the United States. The schedule of Division I games can be represented by a network, in which the nodes denote the 115 teams and the edges represent 613 games played in the course of the year. The teams are divided into 12 conferences containing around 8–12 teams each. Games are more frequent between members of the same conference than those between members of different conferences, with teams playing an average of about seven intraconference games and four interconference games in the 2000 season. Interconference play is not uniformly distributed; teams that are geographically close to one another but belong to different conferences are more likely to play with one another than teams separated by large geographic distances. The natural community structure in the network makes it a commonly used workbench for community-detecting algorithm testing [25,26,28,29].

Using our algorithm, we can partition the network into conferences with a high degree of success. Figure 4 shows

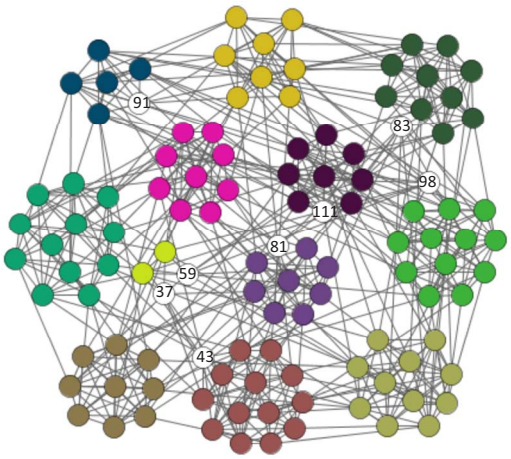

FIG. 4. (Color online) Community structure of the football team network.

the community structure of football team network calculated by our method. Because there are few edges between 5 members of the 12th conference, these 5 nodes are distributed to other communities; e.g., node 91 is distributed to community 9, node 43 is distributed to community 6, node 81 is distributed to community 8, and node 83 is distributed to community 5. We note that nodes 37, 59, 60, and 64 construct to a new community because there are more links within them than with the other nodes. Nodes 98 and 111 are incorrectly classified due to the fact that there are more games with the teams in the classified communities than the teams in their own conferences. The community structure found by our method seems to suggest a more precise organization than the original conferences.

3. Journal index network

The journal index network constructed by Rosvall and Bergstrom [17] consists of 40 journals as nodes from 4 different fields: physics, chemistry, biology, and ecology and 189 links connecting nodes if at least one article from one journal cites an article in the other journal during 2004. Ten journals with the highest impact factor in the 4 different fields were selected. Using our method, we can partition the network into 4 communities correctly (Fig. 5).

We can also partition the network into two, three, or five modules, but such partitions yield lower D values. When we partition the network into two components, physical journals cluster together with chemical journals and biological journals cluster together with ecological journals. When we split it into three components, ecological journals and biological journals separate, but physical journals and chemical journals remain together in a single module. When we intend to split the network into five modules, we get essentially the same partition as with four, only with the singly connected journal Conservation Biology split off by itself as a community. This result is consistent with that in [17].

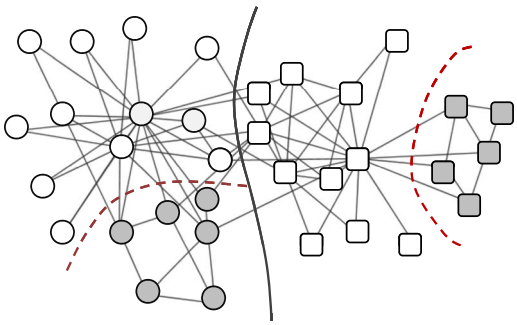

FIG. 3. (Color online) Zachary's karate club network. Square nodes and circle nodes represent the instructor's faction and the administrator's faction, respectively.

VI. CONCLUSION

In this paper, we proposed a measure called modularity density or D value for resolving community structure,

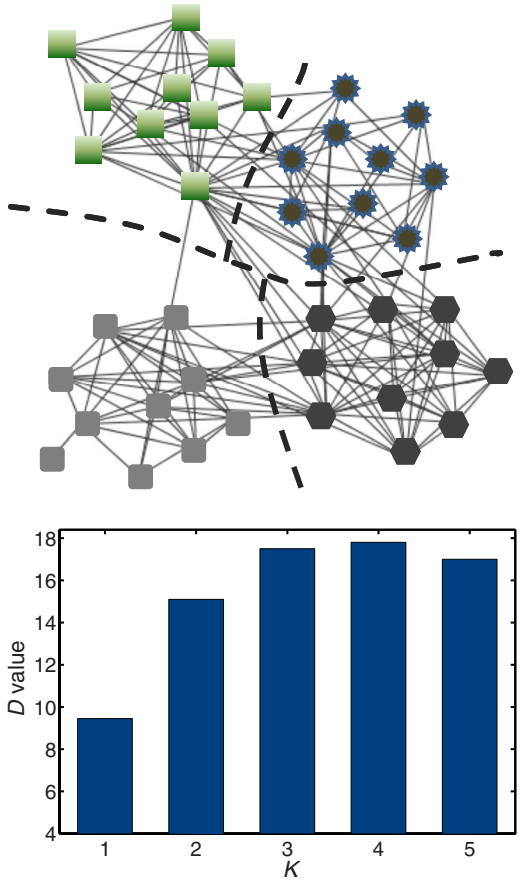

FIG. 5. (Color online) Community structure of the journal index network and the D value for the journal network partition into one to five different modules.

showed that it can be considered as a convex combination of two known indexes and proved that our criterion is equivalent to the objective function of kernel k means. We have verified that optimization of the D value has no drawback to divide the network either into too small communities or into too large communities. By optimizing the D value, we can almost always resolve the network into correct communities.

We also formulated the D-value optimization problem as a nonlinear integer programming (see Appendix A) and conducted numerical tests on both artificial networks and real-world networks. Compared with other algorithms, the D value has no problem in grouping small modules. Our algorithm can generally find the global optimal solution in a short time and also is suited for weighted networks.

By studying the community-detection problem, we may obtain deep insights into the complexity of networks. However, the well-known modularity Q has encountered obvious difficulties and limitations with practical applications [16–18]. From the theoretical and numerical results of this paper, we believe that the proposed measure is a significant contribution to this field. In particular, the general modularity density D_λ of Eq. (10) can be used to resolve various types of communities. The flexible λ also enlarges our understanding about network structures. Moreover, a more efficient optimization technique based on this measure can be expected from the theoretical results of this paper.

ACKNOWLEDGMENTS

This work was partly supported by the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality [PHR(IHLB)] and the Ministry of Science and Technology, China, under Grant No. 2006CB503905, National Natural Science Foundation of China under Grant Nos. 10701080, 10631070, and 60503004, and K.G. Wang Education Foundation Hong Kong. This research work was partly supported by JSPS and NSFC under a JSPS-NSFC collaboration project. The authors thank Professor M. E. J. Newman and Martin Rosvall for providing the network data.

APPENDIX A: NONLINEAR INTEGER PROGRAMMING MODEL FOR OPTIMIZING THE D VALUE

The D-value optimization problem can be formulated as an integer nonlinear programming problem.

Given a network $G=(V,E)$ with $V=\{v_1,\cdots,v_n\}$, the adjacent matrix of G is A:

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix},$$

where

$$a_{ij}=\begin{cases} 1 & \text{if } (v_i,v_j)\in E, \\ 0 & \text{otherwise.} \end{cases}$$

Let x_{il} $(i=1,\ldots,n,\ l=1,\ldots,k)$ be a set of binary variables, where $x_{il}=1$ denotes that the node v_i belongs to the lth community. The problem of dividing network into k communities can be modeled as follows:

$$\max f=\sum_{l=1}^{k}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_{il}x_{jl}-\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_{il}(1-x_{jl})}{\sum_{i=1}^{n}x_{il}}, \quad \text{(A1)}$$

subject to

$$0<\sum_{i=1}^{n}x_{il}<n, \quad l=1,\ldots,k,$$

$$\sum_{l=1}^{k}x_{il}=1, \quad i=1,\ldots,n,$$

$$x_{il}=0,1, \quad i=1,\ldots,n, \quad l=1,\ldots,k,$$

when $k=2$. We can use binary variables x_i $(i=1,\cdots,n)$ to express the division of the network; that is, $x_i=1$ denotes that the node v_i belongs to the first community, while $x_i=0$ denotes that v_i belongs to the second community. So the model can be expressed as follows:

$$\max f = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j - \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i(1-x_j)}{\sum_{i=1}^{n} x_i} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(1-x_i)(1-x_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(1-x_i)x_j}{n - \sum_{i=1}^{n} x_i}, \tag{A2}$$

subject to

$$0 < \sum_{i=1}^{n} x_i < n,$$

$$x_i = 0,1, \quad i = 1, \ldots, n.$$

Although the integer nonlinear programming is theoretically difficult to solve, the constraint conditions in above models are simple. Hence, we can directly solve the relaxed problem with the continuous variables in [0,1]. Experimental results show that almost we can always obtain an integer optimal solution by solving the relaxed problem.

APPENDIX B: COMPARISON OF THE *Q* AND *D* VALUES ON A CLIQUE-RING NETWORK

In this appendix we mathematically show the differences between the modularity Q and the modularity density D. Both the partition quality functions are of complicated behaviors, varying with the objective networks. So some templet networks are chosen for their comparison.

In the analysis, we intend to derive continuous fitting functions $\Phi_Q(x)$ and $\Phi_D(x)$ such that

$$\Phi_Q(K) = \max_{G_1,\ldots,G_K} Q(G_1, \ldots, G_K) = \max_{G_1,\ldots,G_K} \sum_{i=1}^{K} \left[\frac{|E(G_i)|}{|E(G)|} - \left(\frac{2|E(G_i)| + |E(G_i, \bar{G}_i)|}{2|E(G)|} \right)^2 \right] \tag{B1}$$

and

$$\Phi_D(K) = \max_{G_1,\cdots,G_K} D(G_1, \ldots, G_K) = \max_{G_1,\ldots,G_K} \sum_{i=1}^{K} \frac{2|E(G_i)| - |E(G_i, \bar{G}_i)|}{|V(G_i)|}, \tag{B2}$$

where $G_1 \cup G_2 \cdots \cup G_K = G$ and G is the objective network. $\Phi_Q(x)$ and $\Phi_D(x)$ are not always easily derived even for some simple templet networks.

The templet network has n m-cliques as the nodes in the simple ring network. In this case, suppose that we have K communities, each of which consists of n_i cliques,

$$\begin{aligned}
\Phi_{Qcr}(K) &= \max_{G_1,\ldots,G_K} Qcr(G_1, \ldots, G_K) \\
&= \max_{G_1,\ldots,G_K} \sum_{i=1}^{K} \left\{ \frac{n_i m(m-1)/2 + n_i - 1}{nm(m-1)/2 + n} - \left[\frac{n_i m(m-1)/2 + (n_i - 1) + 1}{nm(m-1)/2 + n} \right]^2 \right\} \\
&= \frac{1}{[nm(m-1) + 2n]^2} \max_{G_1,\ldots,G_K} \left\{ n^2[m(m-1)+2]^2 - 2[nm(m-1)+2n]K - \sum_{i=1}^{K} n_i^2[m(m-1)+2]^2 \right\} \\
&= 1 - \frac{2K}{nm(m-1)+2n} - \min_{G_1,\ldots,G_K} \frac{1}{n^2} \sum_{i=1}^{K} n_i^2 \\
&= 1 - \frac{2K}{nm(m-1)+2n} - \frac{1}{K};
\end{aligned} \tag{B3}$$

then,

$$\Phi_{Qcr}(x) = 1 - \frac{2x}{nm(m-1)+2n} - \frac{1}{x} \quad (1 \le x \le n), \tag{B4}$$

$$[\Phi_{Qcr}(x)]' = -\frac{2}{nm(m-1)+2n} + \frac{1}{x^2}, \tag{B5}$$

which implies the optimal solution

$$x^*_{Qcr} = \sqrt{m(m-1)/2 + 1}\sqrt{n} \quad \text{for } m(m-1) \le 2(n-1). \tag{B6}$$

When $m(m-1) > 2(n-1)$, $x*Qcr = n$. When $m=1$, $x^*_{Qcr} = \sqrt{n}$. On the other hand, when applying the modularity density D to the ring of cliques, we analyze $\Phi_{Dcr}(K)$ for $K \le n$ just following the computation of $\Phi_{Dr}(K)$:

$$\begin{aligned}
\Phi_{Dcr}(K) &= \max_{G_1,\ldots,G_K} Drc(G_1, \ldots, G_K) \\
&= \max_{G_1,\ldots,G_K} \sum_{i=1}^{K} \frac{n_i m(m-1) + 2(n_i - 1) - 2}{n_i m} \\
&= \max_{G_1,\ldots,G_K} \sum_{i=1}^{K} \frac{n_i[m(m-1)+2] - 4}{n_i m} \\
&= \frac{m(m-1)+2}{m} K - \frac{4}{m}\frac{K^2}{n} \\
&= \frac{m(m-1)+2}{m} K - \frac{4}{mn} K^2;
\end{aligned} \tag{B7}$$

then

$$\Phi_{Dcr}(x) = \frac{m(m-1)+2}{m} x - \frac{4}{mn}. \tag{B8}$$

To find the optimal partition, we solve the problem

$$\max \Phi_{Dcr}(x) = \frac{m(m-1)+2}{m}x - \frac{4}{mn}x^2 \quad \text{subject to } 1 \le x \le n, \tag{B9}$$

which is a simple linearly constrained convex programming problem. Solving the corresponding Khun-Tucker equation leads to the optimal solution

$$x^*_{Dcr} = \frac{n}{4}, \quad m = 1,$$

$$x^*_{Dcr} = \frac{n}{2}, \quad m = 2,$$

$$x^*_{Dcr} = n, \quad m \ge 3. \tag{B10}$$

Therefore, the community size found by the D value is unrelated to the total size of the network, mn, but the community size found by the modularity Q is related to the total size of the network.

[1] L. C. Freeman, Am. J. Sociol. **98**, 152 (1992).
[2] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003).
[3] S. Zhang, G. Jin, X.-S. Zhang, and L. Chen, Proteomics **7**, 2856 (2007).
[4] G. Jin, S. Zhang, X.-S. Zhang, and L. Chen, PLoS ONE **2**, e1207 (2007).
[5] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).
[6] L. Danon, A. Dìaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech.: Theory Exp. (2005) P09008.
[7] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[8] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[9] S. Zhang, R. S. Wang, and X. S. Zhang, Physica A **374**, 483 (2007).
[10] S. Zhang, R. S. Wang, and X. S. Zhang, Phys. Rev. E **76**, 046103 (2007).
[11] L. Danon, A. Díaz-Guilera, and A. Arenas, J. Stat. Mech.: Theory Exp. (2006) P11010.
[12] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[13] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[14] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[15] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[16] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[17] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. U.S.A. **104**, 7327 (2007).
[18] S. Muff, F. Rao, and A. Caflisch, Phys. Rev. E **72**, 056107 (2005).
[19] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[20] I. S. Dhillon, Y. Guan, and B. Kulis, in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York, 2004), pp. 551–556.
[21] N. Cristianini and J. Shawe-Taylor, *Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods* (Cambridge University Press, Cambridge, England, 2000).
[22] L. Angelini, S. Boccaletti, D. Marinazzo, M. Pellicoro, and S. Stramaglia, Chaos **17**, 023114 (2007).
[23] A. Arenas, A. Fernandez, and S. Gomez, e-print arXiv:physics/0703218.
[24] E. Ravasz, A. L. Somera, D. A. Mongru, A. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).
[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] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[28] H. J. Zhou, Phys. Rev. E **67**, 061901 (2003).
[29] S. Zhang, X. Ning, and X. S. Zhang, Eur. Phys. J. B **57**, 67 (2007).