Eur. Phys. J. B **77**, 547–557 (2010)
DOI: 10.1140/epjb/e2010-00290-3

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Fuzzy modularity and fuzzy community structure in networks

Jian Liu[a]

LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

Received 26 November 2009 / Received in final form 8 June 2010
Published online 1st October 2010 –

Abstract. To find the fuzzy community structure in a complex network, in which each node has a certain probability of belonging to a certain community, is a hard problem and not yet satisfactorily solved over the past years. In this paper, an extension of modularity, the fuzzy modularity is proposed, which can provide a measure of goodness for the fuzzy community structure in networks. The simulated annealing strategy is used to maximize the fuzzy modularity function, associating with an alternating iteration based on our previous work. The proposed algorithm can efficiently identify the probabilities of each node belonging to different communities with random initial fuzzy partition during the cooling process. 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.

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, powergrid, 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 number of communities [10–35], 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 [36].

In a previous paper [30], we have presented a probabilistic framework for network partition, which can be considered as a natural extension of either the fuzzy c-means algorithm in statistics [37] to network partitioning, or the deterministic framework for optimal network partition presented in [29]. In traditional network partitioning problem, each node belongs to only one community after the partition. However, this is often too restrictive, since in many diffusive networks nodes at the boundary among communities share commonalities with more than one community and play a role of transition. This motivates the idea of probabilistic partition [30] to provide a much more mature way of discussing network structure. Instead of assigning nodes to specific communities, we make each node has a certain probability of belonging to a certain community. The methods in [30] perform successfully but need a known number of communities as a model parameter and a deterministic partition detected by an analogy to k-means in [29] as a special initialization.

As described in [30], the best probabilistic partition of a network into communities is such that the Markov chain describing a random walk on the coarse grained graph, whose nodes are the communities of the network, gives the best approximation of the full random walk dynamics on this network. The quality of the approximation is given by a metric between the stochastic matrices of the two processes, which thus needs to be minimized. On a related but quite different front, [28] is also along the lines of random walker Markovian dynamics and makes clusters via reducing the Markov chain. However, network partitioning turns into a coding problem that finding the partition yields the minimum description length of an infinite random walk, with idea of describing a network by using less information than that encoded in the full adjacency matrix. The goal is to optimally compress the information needed to describe the process of information diffusion across the network.

[a] e-mail: dugujian@pku.edu.cn

However, people are sometimes required to determine the number of communities in a network and meet the difficulty that the objective function in [30] decreases as the number of communities increases. The initial membership, set by a special indicator matrix detected by k-means [29], also costs extremely much. To overcome these weaknesses, we develop an effective approach to finish automatic model selection in the current paper. As is well known, a concept of modularity has been broadly used to evaluate the goodness of the network communities [11–16,21,22], which has larger values indicating stronger community structure. It can be naturally extended to a fuzzy version in order to quantify the quality of a probabilistic partition and thus help in the process of automatic model selection by not requiring users to fix the number of communities. Then simulated annealing [22,38,39] and corresponding strategies are proposed to obtain the maximal value of this fuzzy modularity function. The cooling process is operated with an alternating iterative procedure based on our previous work [30]. We should remark that the present method can avoid ineffectively repetition and lead to a high degree of efficiency and accuracy, since the iteration accelerates the tendency of maximizing the fuzzy modularity function, and a larger value of fuzzy modularity can be reached than searching over all possible N by making use of the method in [30] for each fixed N(see Fig. 4). In general, this approach can not only identify the probabilities of each node belonging to different communities, but also determine the optimal number of communities automatically without any prior knowledge about the community structure. In addition, the initial values of the memberships can also be randomly chosen. The fuzzy community structure contains more detailed information and has more predictive power than the old way of doing network partition. How to draw lesson from the validity index [40] in traditional clustering literature, making use of a variation of the optimal prediction error involves both compactness and separation between the groups, to provide a better quantity to measure the partition, will be our future work.

We will construct our algorithm – Simulated Annealing to maximize the Fuzzy Modularity (SAFM) associating with an alternating iteration based on the Euler-Lagrange equations in [30]. The algorithm is tested on some artificial networks, including the ad hoc network with 128 nodes, the LFR benchmark for normal network and network with overlapping communities. The numerical results suggest that our algorithm is efficiently implemented with reasonable computational effort and leads to an accurate probabilistic partition. Moreover, successful application to several real-world networks, including the karate club network, the dolphins network, the American political books 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 probabilistic framework for network partition in [30] and then propose the definition of the fuzzy modularity function. The algorithm (SAFM) and the corresponding strategies are described in detail in Section 3. In Section 4, we apply the proposed algorithm to the representative examples mentioned before. Finally we make the conclusion in Section 5.

2 Framework for fuzzy community structure in networks

2.1 The probabilistic partition of networks

We will start with a brief review of the probabilistic framework for network partition proposed in [30]. Let $G(S,E)$ be a network with n nodes, 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 sum over weights on all the edges is denoted by m and naturally $m=\sum_{x,y\in S}e(x,y)/2$. 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 $G(S,E)$ [9,18,19,28–30,41]. 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 [42,43]. This defines the transition matrix $p=(p(x,y))$ of random walk processes. Here we restrict ourselves to an undirected network, i.e. $e(x,y)=e(y,x)$, and if we define

$$\mu(x)=\frac{d(x)}{\sum_{z\in S}d(z)} \tag{2}$$

then μ is a stationary distribution of this Markov chain and it satisfies the detailed balance condition $\mu(x)p(x,y)=\mu(y)p(y,x)$ [9,29,30,41].

The basic idea in [30] is to introduce a metric for the stochastic matrix $p(x,y)$

$$\|p\|_\mu^2=\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p(x,y)|^2 \tag{3}$$

and find the reduced Markov chain $\tilde{p}$ by minimizing the distance $\|\tilde{p}-p\|_\mu$.

For a given partition of S as $S=\cup_{k=1}^N S_k$ with $S_k\cap S_l=\emptyset$ if $k\neq l$, let $\hat{p}_{kl}$ be the coarse grained transition probability from S_k to S_l on the state space $\mathbb{S}=\{S_1,\ldots,S_N\}$ which satisfies

$$\hat{p}_{kl}\geq 0\quad\text{and}\quad\sum_{l=1}^N\hat{p}_{kl}=1, \tag{4}$$

and let $\rho_k(x)$ be the probability of the node x belonging to the kth community which needs the assumption that

$$\rho_k(x) \geq 0 \quad \text{and} \quad \sum_{k=1}^{N} \rho_k(x) = 1, \quad \text{for all} \quad x \in S. \tag{5}$$

Naturally the matrix $\tilde{p}$ can be lifted to the space of stochastic matrices on the original state space S via

$$\tilde{p}(x,y) = \sum_{k,l=1}^{N} \rho_k(x)\hat{p}_{kl}\mu_l(y), \qquad x,y \in S, \tag{6}$$

where

$$\mu_k(x) = \frac{\rho_k(x)\mu(x)}{\hat{\mu}_k} \quad \text{and} \quad \hat{\mu}_k = \sum_{z\in S} \rho_k(z)\mu(z). \tag{7}$$

The idea of lifting the size of stochastic matrices expresses the perspective that the node x transits to y through different channels from community k to community l with their corresponding belonging probabilities and stay there in equilibrium state. It is not difficult to show that $\tilde{p}$ is indeed a transition probability matrix and satisfies the detailed balance condition with respect to μ if $\hat{p}_{kl}$ satisfies the detailed balance condition with respect to $\hat{\mu}$.

Given the number of the communities N, we optimally reduce the random walker dynamics by considering the following minimization problem

$$\min_{\rho_k(x),\ \hat{p}_{kl}} J = \|p - \tilde{p}\|_\mu^2 = \sum_{x,y\in S} \mu(x)\mu(y) \times \left| \sum_{m,n=1}^{N} \rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n} - \frac{p(x,y)}{\mu(y)} \right|^2, \tag{8}$$

subject to the constraints (4) and (5). To minimize the objective function J in (8), we define

$$\begin{aligned} \hat{p}^*_{kl} &= \sum_{x,y\in S} \mu_k(x)p(x,y)\rho_l(y) \\ &= \frac{1}{\hat{\mu}_k} \sum_{x,y\in S} \mu(x)\rho_k(x)p(x,y)\rho_l(y). \end{aligned} \tag{9}$$

Then $\hat{p}^*_{kl}$ is indeed a stochastic matrix and satisfies the detailed balance condition with respect to $\hat{\mu}$.

The optimization of J with constraints $\sum_{k=1}^{N} \rho_k(x) = 1$ corresponds to find the critical points of (8). The Euler-Lagrange equations are derived as

$$\left(I_{\hat{\mu}}^{-1} \cdot \hat{\mu}\right) \cdot \hat{p} \cdot \left(I_{\hat{\mu}}^{-1} \cdot \hat{\mu}\right) = \hat{p}^*, \tag{10a}$$

$$\rho = I_{\hat{\mu}}\hat{p}^{-1}\hat{\mu}^{-1}\rho p^T, \tag{10b}$$

where $\rho = (\rho_k(x))_{k=1,\ldots,N,x\in S}$ is a $N \times n$ matrix and $\hat{\mu}$ is a $N \times N$ matrix with entries

$$\hat{\mu}_{kl} = \sum_{z\in S} \mu(z)\rho_k(z)\rho_l(z) = (\rho \cdot I_\mu \cdot \rho^T)_{kl}. \tag{11}$$

The diagonal matrices I_μ, $I_{\hat{\mu}}$ are $n \times n$ and $N \times N$ respectively, with entries

$$I_\mu(x,y) = \mu(x)\delta(x,y), \quad x,y \in S, \tag{12a}$$

$$(I_{\hat{\mu}})_{kl} = \hat{\mu}_k\delta_{kl}, \quad k,l = 1,\ldots,N, \tag{12b}$$

where $\delta(x,y)$ and δ_{kl} are both Kronecker delta symbols.

A strategy suggested immediately by the Euler-Lagrange equations (10) is to iterate alternatively between the equations for $\hat{p}$ and ρ. To ensure the nonnegativity and normalization conditions for $\hat{p}$ and ρ, we add a projection step after each iteration and change (10) to

$$\hat{p} = \mathcal{P}\left(\hat{\mu}^{-1} \cdot I_{\hat{\mu}} \cdot \hat{p}^* \cdot \hat{\mu}^{-1} \cdot I_{\hat{\mu}}\right), \tag{13a}$$

$$\rho = \mathcal{P}\left(I_{\hat{\mu}}\hat{p}^{-1}\hat{\mu}^{-1}\rho p^T\right). \tag{13b}$$

Here $\mathcal{P}$ is a projection operator which maps a real vector into a vector with nonnegative, normalized components.

2.2 Fuzzy modularity

In recent years, a concept of modularity proposed by Newman [11–16,25,27,28] 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

$$\begin{aligned} Q = &\,(\text{number of edges within communities}) \\ &-(\text{expected number of such edges}). \end{aligned}$$

It is a function of the particular partition of the network into groups, with larger values indicating stronger community structure [11–14]. 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 [11–16,21,24,34].

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 $p^E(x,y)$ for an edge to fall between every pair of node x and y is specified. More precisely, $p^E(x,y)$ 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

$$Q = \frac{1}{2m} \sum_{k=1}^{N} \sum_{x,y\in S_k} \left(e(x,y) - p^E(x,y)\right), \tag{14}$$

where

$$p^E(x,y) = \frac{d(x)d(y)}{2m} \tag{15}$$

and m is again 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 literature [15,43].

In order to evaluate the goodness of the fuzzy clustering on different partitions of a network, we extend the modularity (14) to a fuzzy formulation. According to the above framework, an edge connected pair of nodes no longer belongs to a certain community but to different communities with nonzero probabilities, due to the weights each node has of belonging to these groups. This motivates the definition of the fuzzy modularity

$$\begin{aligned} Q_f = &(\text{probabilistic number of edges within communities}) \\ &-(\text{expected probabilistic number of such edges}). \end{aligned}$$

It can be viewed as a criterion to quantify the quality of a probabilistic partition and thus help in the process of automatic model selection by not requiring a known number of communities.

For a particular fuzzy partition $\{\rho_k(x)\}_{k=1}^N$, we classify the nodes according to the majority rule, i.e. if $k = \arg\max_l \rho_l(x)$ for a given node x then we set $x \in S_k$. Then the fuzzy modularity Q_f can also be developed based on the null model [15] via

$$Q_f = \frac{1}{2m}\sum_{k=1}^{N}\sum_{x,y\in S_k}\Big(\frac{\rho_k(x)+\rho_k(y)}{2}e(x,y) - p_f^E(x,y)\Big), \tag{16}$$

where $p_f^E(x,y)$ is expected probabilistic number of edge $e(x,y)$ with the form

$$p_f^E(x,y) = \frac{d_f(x)d_f(y)}{2m}, \quad x,y \in S_k \tag{17}$$

and $d_f(x)$ is the extended degree of node x in community S_k under the probabilistic setting and given by

$$\begin{aligned} d_f(x) = &\sum_{z\in S_k}\frac{\rho_k(x)+\rho_k(z)}{2}e(x,z) \\ &+\sum_{z\notin S_k}\frac{\rho_k(x)+(1-\rho_k(z))}{2}e(x,z). \end{aligned} \tag{18}$$

The extended version (16) can be considered as a generalization of the traditional modularity (14). An ideal partition for a fixed community number N requires a more stable state in $\{\rho_k(x)\}_{k=1}^N$. Thus, an optimal partition can be found by solving

$$\max_N\left\{\max_{\{\rho_k(x)\}_{k=1}^N} Q_f\right\}. \tag{19}$$

We should remark that after deriving a fuzzy value for node membership in communities, the fuzzy modularity is quantified only using pairs of nodes whose maximum memberships are in the same community. This seems to throw away other information in the membership ρ, not just the maximal component. However, the experimental results indicate that the practical utilization of (16) makes different partition from using the original modularity (see Fig. 4) and the rest information of ρ are taken into account as one produces the overlapping communities according to our algorithm (see Fig. 3).

3 The algorithm

The first simulated annealing algorithm was motivated by simulating the physical process of annealing solids [38] and later widely used to optimization problems [39]. Modularity optimization methods [11–13,15,21,22] was often carried out through simulated annealing [22], which is not a fast technique but yields good estimates of modularity maxima. Here our work is quite different from the previous contributions, since the process of iteration accelerates the tendency of maximizing the fuzzy modularity function and avoids ineffectively repetition.

Let $E = -Q_f$. $E^{(n)}$and $E^{(n+1)}$ represent the current energy and new energy respectively. $E^{(n+1)}$ is always accepted if it satisfies $E^{(n+1)} < E^{(n)}$, but if $E^{(n+1)} > E^{(n)}$ the new energy level is only accepted with a probability as specified by $\exp(-\frac{1}{T}\triangle E^{(n)})$, where $\triangle E^{(n)} = E^{(n+1)} - E^{(n)}$ is the difference of energy and T is the current temperature. The initial state $\{\rho_k^{(0)}(x)\}_{k=1}^N$ is generated at random and N is a integer within the range $[N_{\min}, N_{\max}]$. The initial temperature T is set to a high temperature $T_{\max}$. A neighbor of the current state is produced by randomly choosing the strategies of our proposal, then the energy of the new state is calculated. The new state is kept if the acceptance requirement is satisfied. This process will be repeated for R times at the given temperature. A cooling rate $0 < \alpha < 1$ decreased the current temperature until reached the bound $T_{\min}$. The whole procedure of the Simulated Annealing to maximize the Fuzzy Modularity (SAFM) associating with an alternating iteration algorithm is summarized below

(1) Set parameters $T_{\max}$, $T_{\min}$, $N_{\min}$, $N_{\max}$, α and R. Choose N randomly within range $[N_{\min}, N_{\max}]$ and initialize the memberships $\{\rho_k^{(0)}\}_{k=1}^N$ randomly; Set the current temperature $T = T_{\max}$.
(2) Compute $\hat{p}^{(0)}$ according to (13a) and calculate the initial energy $E^{(0)}$ using (16); Set $n^* = 0$.
(3) For $n = 0, 1, \ldots, R$, do the following
  (3.1) generate a set of new memberships $\{\rho_k^{(n)}\}_{k=1}^{N'}$ according to our proposal below and set $N = N'$;
  (3.2) update $\hat{p}^{(n+1)}$, $\rho^{(n+1)}$ and the new energy $E^{(n+1)}$ according to (13a), (13b) and (16);
  (3.3) accept or reject the new state. If $E^{(n+1)} < E^{(n)}$ or $E^{(n+1)} > E^{(n)}$ with $u \sim \mathcal{U}[0,1]$, $u < \exp\{-\frac{1}{T}\triangle E^{(n)}\}$, accept the new solution, i.e. set $n = n+1$; else, reject it;
  (3.4) update the optimal state. If $E^{(n)} < E^{(n^*)}$, set $n^* = n$.
(4) Cooling temperature: $T = \alpha \cdot T$. If $T < T_{\min}$, go to Step(5); else, set $n = n^*$, repeat Step(3).

(5) Output the optimal solution $\{\rho_k^{(n^*)}\}_{k=1}^N$ and the maximum fuzzy modularity $Q_f = E^{(n^*)}$ of the whole procedure. The majority rule

$$S_k = \{x : k = \arg\max_l \rho_l^{(n^*)}(x)\} \tag{20}$$

gives the deterministic partition and

$$S_k = \{x : \rho_k^{(n^*)}(x) > \eta\} \tag{21}$$

produces the overlapping communities, where $\eta > 0$ is a threshold.

The choice of the projection operator $\mathcal{P}$ used in our computation is the direct projection to the boundary. Let $\boldsymbol{v} = (v_1, v_2, \ldots, v_N) \in \mathbb{R}^N$ and $\Lambda = \{i : v_i \geq 0\}$. When $i \notin \Lambda$, we set $\mathcal{P}v_i = 0$; otherwise we set $\mathcal{P}v_i = v_i / \sum_{j\in\Lambda} v_j$.

Our proposal to the process of generating a set of new memberships in step (3.1) comprises three functions, which are deleting a current community, splitting a current community and keeping a current community. At each iteration, one of the three functions can be randomly chosen and the community size

$$M_k = \sum_{x\in S} \rho_k(x), \quad k = 1, \ldots, N, \tag{22}$$

is used to select a community. Obviously, the existence of a community is more probable if its size is larger. The three functions are described below

(i) Delete Community. The community with the minimal community size M_d is identified, which should be deleted the dth row from the current membership matrix ρ and added to $\rho_k = \rho_k + \rho_d$, where $k = \arg\max_m \hat{p}^*_{dm}$ and $\hat{p}^*$ is defined in (9).
(ii) Split Community. The community with the maximal community size M_s is chosen, which should be replaced by two new communities. We take $r(x) \sim^{i.i.d.} \mathcal{U}[0,1]$, then $\rho_{N+1}(x) = r(x) \cdot \rho_s(x), \rho_s(x) = (1 - r(x)) \cdot \rho_s(x), \forall x \in S$.
(iii) Keep Community. We remain the current communities.

The number of the iteration steps depends on the initial and terminal temperature and the cooling rate. For each iteration, the total cost in in the step of computing $\hat{p}$ is $O(N^2(m+n))$ and the cost for ρ is $O(N^2 n + Nm)$. The global maximal problem (19) can be also solved by searching over all the possible N, and for each fixed N using the alternating iteration (13). But this will cost extremely much since the initialization of k-means in [29] should be implemented approximate 2000 trials due to its local minima for each parameter N. However, SAFM can avoid repeating ineffectively and lead to a higher degree of efficiency and accuracy. Moreover, a more optimal partitioning result with a larger value of Q_f can be obtained than searching over all possible N by making use of (13) for each fixed N(see Fig. 4).

Another advantage of SAFM is that it overcomes the weaknesses of the methods in [30]. The cooling process can efficiently and automatically determine the number of communities N without fixing it as a known model parameter, and the initial memberships $\{\rho_k^{(0)}\}$ can be randomly chosen, rather than taken as the indicator matrix for each node detected by the k-means in [29].

4 Experimental results

4.1 Artificial networks

4.1.1 The ad hoc network with 128 nodes

The first example is the ad hoc network with 128 nodes. The ad hoc network is a benchmark problem used in many papers [10,11,21,24,26,29,30]. It has a known community structure and is constructed as follows. Suppose we choose $n = 128$ nodes, 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

$$31p_{\text{in}} + 96p_{\text{out}} = 16. \tag{23}$$

We will denote $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. The parameters are set by $T_{\max} = 3.0$, $T_{\min} = 0.01$, $N_{\max} = n/2$, $N_{\min} = 2$, $\alpha = 0.9$ and $R = 50$ in this model computation. If we partition the network with the majority rule after applying our algorithm, that is, classify the nodes according to their maximal weight, we obtain a deterministic partition. To compare the built-in modular structure with the one delivered by the algorithm, we adopt the normalized mutual information, a measure of similarity of partitions borrowed from information theory, which has proved to be reliable [24,32–34]. We change $z_{\text{out}} = 96p_{\text{out}}$ from 0.5 to 10 and look into the normalized mutual information. As we can see from Figure 1, the normalized mutual information varies in a similar way across the different methods as z_{out} increases and the communities become more diffuse at the same time. The natural partition is always found up until $z_{\text{out}} = 6$, then the method starts to fail. It seems that SAFM performs competitive with Newman's fast algorithm [13] and the extremal optimization algorithm [21], especially for the more complicated cases when z_{out} is higher. This also verifies the accuracy of our method while our method gives more detailed information for each node.

4.1.2 The LFR benchmark

The LFR benchmark [32–34] 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

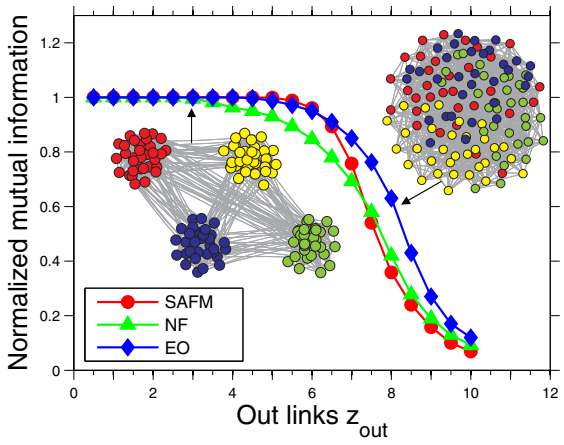

Fig. 1. (Color online) Test of our algorithm compared with Newman's fast algorithm [13] and the extremal optimization algorithm [21] on the ad hoc network with 128 nodes [24,34]. The Ad hoc networks has four communities: For lower $z_{\rm out}$ the communities are easily distinguished, while for higher $z_{\rm out}$ this becomes more complicated.

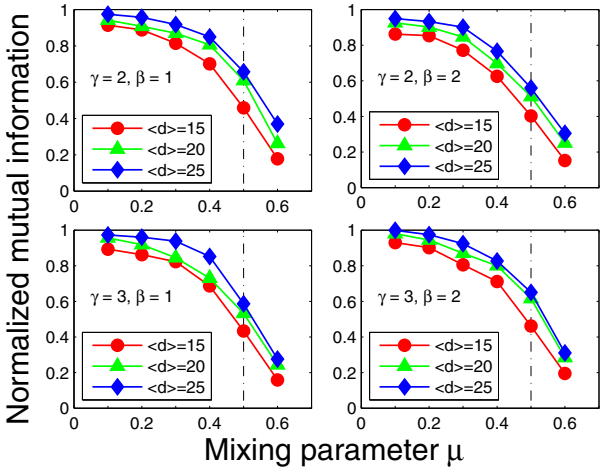

Fig. 2. (Color online) Test of our algorithm on the LFR benchmark [32]. The number of nodes $n = 500$. The results clearly depend on all parameters of the benchmark, from the exponents γ and β to the average degree $\langle d\rangle$. The threshold $\mu_c = 0.5$, shown by dashed vertical line in the 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. Each point corresponds to an average over 20 graph realizations.

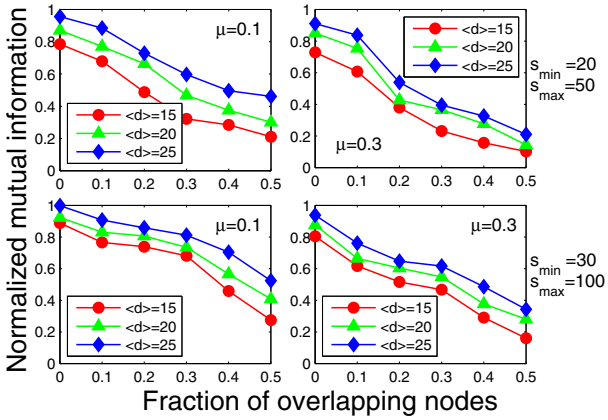

Fig. 3. (Color online) Test of our algorithm on the FLR benchmark for undirected and unweighted networks with overlapping communities [33,34]. The plot shows the variation of the normalized mutual information between the planted and the recovered partition, in its generalized form for overlapping communities [31], with the fraction of overlapping nodes. The networks have 500 nodes, the other parameters are $\gamma = 2, \beta = 1$ and $d_{\rm max} = 50$. Each point corresponds to an average over 20 graph realizations.

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 [32]. The LFR benchmark is further extended to the overlapping communities [33,34] and the corresponding normalized mutual information in its generalized form for overlapping communities is developed in order to test algorithms in [31].

In Figure 2, we show what happens if our algorithm is implemented on the benchmark with $n = 500$. The parameters are set by $T_{\rm max} = 3.0$, $T_{\rm min} = 0.01$, $N_{\rm max} = 50$, $N_{\rm min} = 2$, $\alpha = 0.9$ and $R = 20$ in this model computation. 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' 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 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.

In Figure 3, we show the results of tests performed with our method on the benchmarks with overlapping communities [33,34]. The networks have 500 nodes, the other parameters are $\gamma = 2, \beta = 1$ and $d_{\rm max} = 50$. In this case, the mixing parameter μ is fixed and one varies the fraction of overlapping nodes between communities. We look into the variation of the normalized mutual information between the planted and the recovered partition, in its generalized form for overlapping communities [31] for different average degree $\langle d\rangle$, by setting the same parameters in computation as in Figure 2. Here $\eta = 0.3$ is chosen to produce the overlapping communities. We also notice that the performance of the method is better the larger the average degree $\langle d\rangle$, whereas it gets worse when the fraction of overlapping nodes became larger. In the two top diagrams community sizes range between $s_{\rm min} = 20$ and $s_{\rm max} = 50$, whereas in the bottom diagrams the range

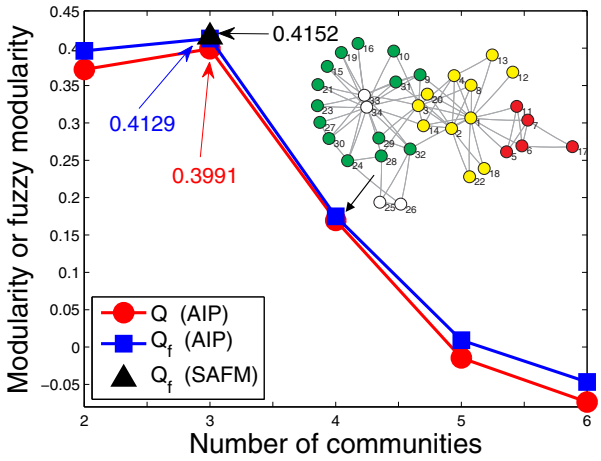

Fig. 4. (Color online) The original modularity and fuzzy modularity functions detected using the method AIP in [30]. It shows clearly that SAFM can find a larger value of $Q_f = 0.4152$ than searching over N using the corresponding AIP in [30]. The insert graph represent the community structure obtained by using AIP with $N = 4$ and the partitioning results become complicated as $N \geq 4$.

goes from $s_{\min} = 30$ and $s_{\max} = 100$. By comparing the diagrams in the top with those in the bottom we see that the algorithm performs better when communities are larger on average.

4.2 Real-world networks

4.2.1 The karate club network

This network was constructed by Wayne Zachary after he observed social interactions between members of a karate club at an American university [44]. Soon after, a dispute arose between the clubs administrator and main teacher and the club split into two smaller clubs. It has been used widely to test the algorithms for finding communities in networks [10,11,14–16,21,26,29–31].

We implement the method by setting $T_{\max} = 3.0$, $T_{\min} = 10^{-5}$, $N_{\max} = 10$, $N_{\min} = 2$, $\alpha = 0.9$ and $R = 50$ and confirm that the simulated annealing approach the iteration can reach a larger fuzzy modularity $Q_f = 0.4152$ than searching over N by using the method AIP in [30] for each fixed N, which are illustrated in detail in Figure 4. The numerical results obtained by our method are presented in Table 1. Figure 5a shows the 3 communities represented by different colors obtained by the majority rule. However, we have more detailed information in fact. From Table 1, we find $\rho_R = 1$ for nodes {5, 6, 7, 11}, $\rho_Y = 1$ for nodes {2, 4, 8} and $\rho_G = 1$ for nodes {15, 16, 19, 21, 23, 27} which mostly lie at the boundary of the green colored group. The others belong to the three colored groups with nonzero probability, especially the nodes {1, 3, 9, 10, 12, 20, 29, 31} have more diffusive weights and they play the role of transition nodes among the corresponding groups. The fuzzy community structure visualized by the weighted average are shown in Figure 5b.

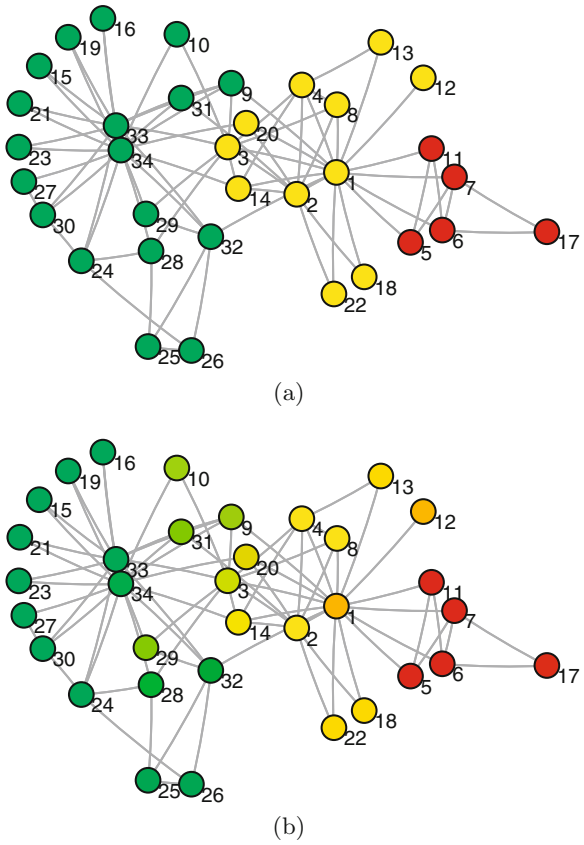

Fig. 5. (Color online) (a) The community structure of the karate club network detected by our algorithm according to the node's maximal weight, with 3 communities represented by different colors. (b) The fuzzy community structure visualized by the weights $\{\rho_k(x)\}$, corresponding to the fuzzy modularity $Q_f = 0.4152$.

Assume that the vectorial representations for different colors in the visualization tool are $C_k, k = 1, \ldots, N$. Then the color vector for the node x is given by the weighted average

$$c(x) = \sum_{k=1}^{N} \rho_k(x) C_k, \quad x \in S. \tag{24}$$

Here the vectorial representations for the colors red, yellow and green in the visualization tool are C_R, C_Y and C_G respectively, and the color vector for the node x is given by $\rho_R(x)C_R + \rho_Y(x)C_Y + \rho_G(x)C_G$. This shows more clearly the transition between different communities. One would naturally speculate that the members in the middle are somewhat closely associated with these communities.

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 [45,46]. The network was compiled from the studies of the dolphins, with ties between dolphin pairs being established

Table 1. The association probability of each node belonging to different communities of the karate club network. ρ_R, ρ_Y or ρ_G means the probability belonging to red, yellow or green colored community in Figure 5, respectively.

Nodes	1	2	3	4	5	6	7	8	9	10	11	12
ρ_R	0.3322	0	0	0	1.0000	1.0000	1.0000	0	0.0293	0	1.0000	0.3165
ρ_Y	0.6678	1.0000	0.6841	1.0000	0	0	0	1.0000	0.4598	0.4920	0	0.6835
ρ_G	0	0	0.3159	0	0	0	0	0	0.5109	0.5080	0	0
Nodes	13	14	15	16	17	18	19	20	21	22	23	24
ρ_R	0.0780	0	0	0	0.9741	0.0780	0	0.0587	0	0.0780	0	0.0105
ρ_Y	0.9220	0.9482	0	0	0	0.9220	0	0.7695	0	0.9220	0	0
ρ_G	0	0.0518	1.0000	1.0000	0.0259	0	1.0000	0.1718	1.0000	0	1.0000	0.9895
Nodes	25	26	27	28	29	30	31	32	33	34		
ρ_R	0.0340	0.0322	0	0	0	0.0009	0	0.0753	0	0		
ρ_Y	0	0	0	0.1615	0.4170	0	0.4074	0.0653	0.0289	0.0801		
ρ_G	0.9660	0.9678	1.0000	0.8385	0.5830	0.9991	0.5926	0.8594	0.9711	0.9199		

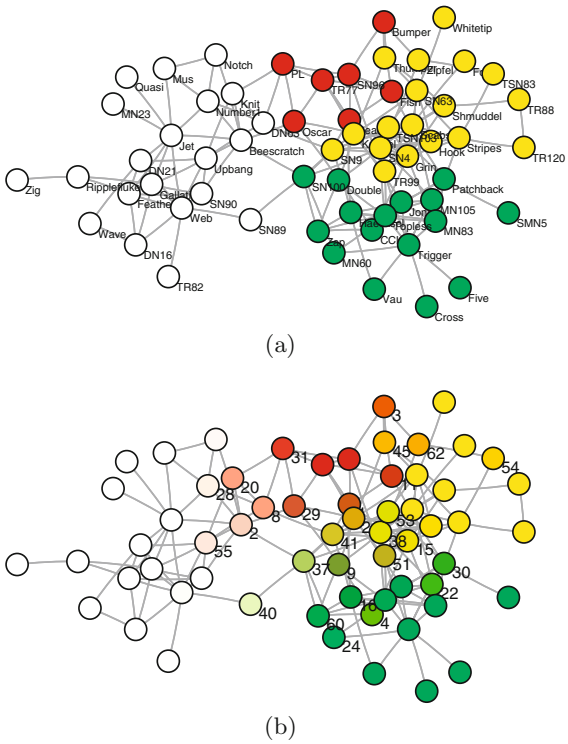

Fig. 6. (Color online) (a) The community structure of the dolphins network detected by our algorithm according to the node's maximal weight, with 4 communities represented by different colors. (b) The fuzzy community structure visualized by the weights $\{\rho_k(x)\}$, corresponding to $Q_f = 0.5050$.

by observation of statistically significant frequent association [11,14,15,31]. The partitioning results are shown in Table 2 and Figure 6. According to Figure 6a, the network seems splitting into two large communities by the white part and the larger one and the larger one keeps splitting into a few smaller communities, represent by different colors. The split into two groups appears to correspond to a known division according to the dolphins' age [46]. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals that the yellow part consists almost of entirely of females and the others almost entirely of males [46]. Table 2 list the association probability of nodes with intermediate weights. For other nodes, though they have not 0–1 weights, one dominate component have strength weight more that 0.95. It seems that nodes {2, 8, 20, 29, 31, 37, 40} take on the role of major connection between the two large groups, and the other nodes in this table play the role of transition among the four communities detected. The visualization of the weights $\{\rho_k(x)\}$ are shown in Figure 6b. In general, the fuzzy partition produced by our method again reflects the extent of frequent interaction in this social network between dolphins.

4.2.3 The American 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 [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 7, nodes have been given whether they are conservative(box) or liberal(diamond), except for a small number of books which are neutral(ellipse). The partitioning results are shown in Figure 7. We find 4 communities denoted by different colors. 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 two remaining community. Thus these books appear to form communities of copurchasing that align closely with political views. Moreover, the fuzzy community structure indicates the tendentiousness of each book on the political views. One could imagine,

Table 2. The association probability of nodes with intermediate weights belonging to different communities of the dolphins network. ρ_R, ρ_Y, ρ_G or ρ_W means the probability belonging to red, yellow, green or white colored community in Figure 6, respectively. For other nodes, though they have not 0–1 weights, one dominate component have strength weight more that 0.95.

Nodes	1	2	3	4	8	9	11	15	16	20	21	22	24	28
ρ_R	0.7389	0.2364	0.7398	0.0216	0.4818	0.2539	0.8918	0.0209	0.1124	0.4829	0.3320	0	0	0.0652
ρ_Y	0.1092	0	0.2602	0.3188	0.0163	0.1229	0	0.8763	0	0	0.5252	0.2589	0	0.0412
ρ_G	0.1519	0.0277	0	0.6596	0	0.6232	0.1082	0.1028	0.8876	0	0.1427	0.7411	0.9445	0
ρ_W	0	0.7359	0	0	0.5019	0	0	0	0	0.5171	0	0	0.0555	0.8936
Nodes	29	30	31	37	38	40	41	45	51	53	54	55	60	62
ρ_R	0.7688	0.0648	0.9053	0.0540	0	0	0.1309	0.3057	0.2103	0	0.1360	0.1341	0.0406	0.3663
ρ_Y	0.0027	0.1602	0	0.3176	0.8352	0.1972	0.6239	0.6943	0.4928	0.7828	0.8640	0.0027	0.0218	0.6337
ρ_G	0.1046	0.7750	0	0.4156	0.1648	0.1836	0.1945	0	0.2969	0.2172	0	0	0.9125	0
ρ_W	0.1239	0	0.0947	0.2128	0	0.6192	0.0507	0	0	0	0	0.8632	0.0251	0

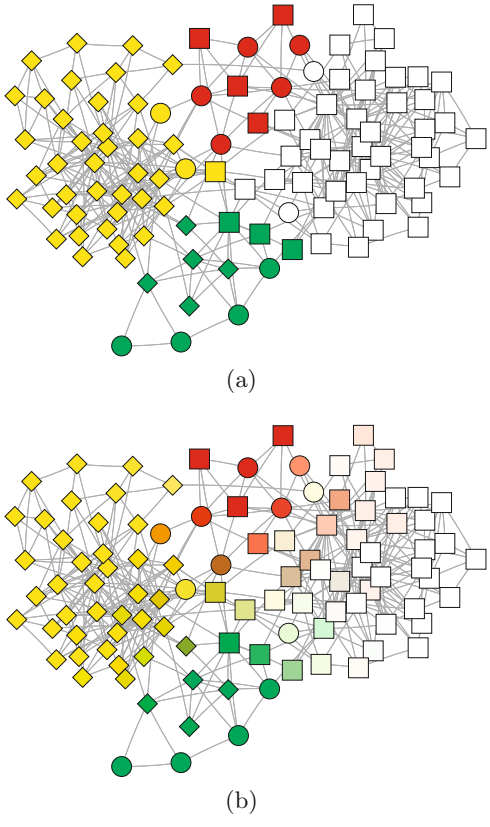

Fig. 7. (Color online) (a) The community structure of the American political books network detected by our algorithm according to the node's maximal weight, with 4 communities represented by different colors. (b) The fuzzy community structure visualized by the weights $\{\rho_k(x)\}$, corresponding to $Q_f = 0.5184$.

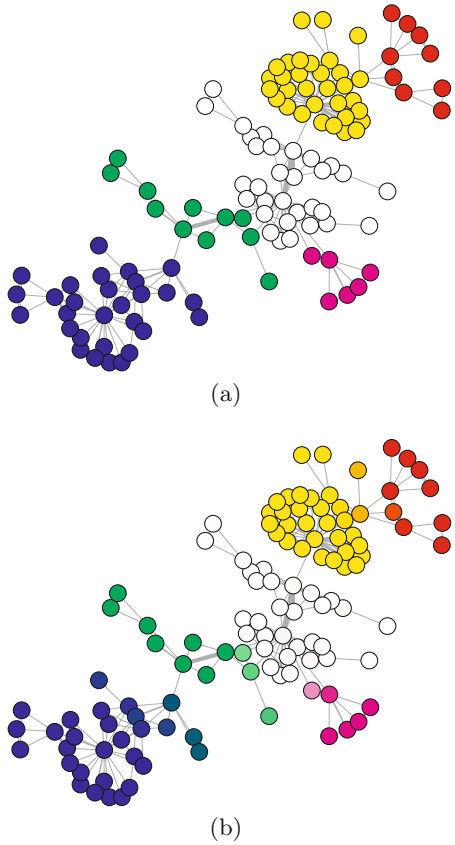

Fig. 8. (Color online) (a) The community structure of the SFI collaboration network detected by our algorithm according to the node's maximal weight, with 6 communities represented by different colors. (b) The fuzzy community structure visualized by the weights $\{\rho_k(x)\}$, corresponding to $Q_f = 0.7075$.

for example, using the algorithm discussed here on the voting recording of the U.S. senators, and predict who is most likely to switch parties.

4.2.4 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 [10,19,20,26]. 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 8a, we illustrate the results from the application of our algorithm to the largest component of the collaboration graph, which consists of 118 scientists. We find that our method split the network into 6 communities with the divisions running principally along disciplinary lines. The community at the top of the figure (red) represents a group of scientists using agent-based models to study problems in economics and traffic flow. The next community (yellow) represents a group of scientists working on mathematical models in ecology, and forms a fairly cohesive structure The largest community (white, magenta, green) is a group working primarily in statistical physics, and seems sub-divided into 3 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. The visualization of weights $\{\rho_k(x)\}$ in Figure 8b shows clearly the extent of cooperation between scientist at the Santa Fe Institute and indicate which research field they tend to join in, with different probabilities. The members who devote themselves to interdisciplinary fields can also be found and evaluated.

5 Conclusions

In this paper, we have proposed the fuzzy modularity function to evaluate the fuzzy community structure in networks. The proposed algorithm – simulated annealing to maximize the fuzzy modularity (SAFM), associating with an alternating iteration, is constructed and succeeds in several representative networks. The experiments show very satisfactory results that our algorithm can identify the probabilities of each node belonging to different communities with a high degree of efficiency and accuracy. The process of simulated annealing with iteration avoid repeating ineffectively and obtain a larger value of Q_f than searching over all possible N by making use of the method in [30] for each fixed N. The present method also successfully overcomes the shortcomings in [30], while the number of communities N can be automatically determined without fixing it as a known model parameter, and the initial memberships can be randomly chosen, rather than taken as the indicator matrix for each node detected by the k-means in [29] at this time. The deterministic partition and the overlapping communities can be produced according to the majority rule and the thresholding operation, respectively. In general, the fuzzy community structure in networks contains more detailed information and has more predictive power than the old way of doing network partition.

The optimal prediction error J in (8) is decreasing as the number of the communities N increases [30], which means the model selection strategy is not self-contained with this proposal. How to draw lesson from the validity index [40] in traditional clustering literature to networks, using a variation of the optimal prediction error instead of the concept of modularity, to obtain a better quantity to evaluate the partition, will be our next step. But the algorithm considered in this paper is efficient and deserved to be investigated.

We thank Professor S. Fortunato for kindly sharing the codes of generating the LFR benchmarks and computing the generalized normalized mutual information. We are also grateful to Professor M.E.J. Newman, Professor H. Zhou and Professor J. Zhang for providing the data of the karate club network, the dolphins network, the political books network and 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. R. Albert, A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
2. M.E.J. Newman, SIAM Rev. **45**, 167 (2003)
3. M.E.J. Newman, A.L. Barabási, D.J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2005)
4. A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek, Physica A **311**, 590 (2002)
5. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabási, Science **297**, 1551 (2002)
6. G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, IEEE Comput. **35**, 66 (2002)
7. J. Shi, J. Malik, IEEE Trans. Pattern Anal. Mach. Intel. **22**, 888 (2000)
8. M. Meilă, J. Shi, A random walks view of spectral segmentation, in *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics* (Kaufmann, San Francisco, 2001), pp. 92–97
9. S. Lafon, A.B. Lee, IEEE Trans. Pattern Anal. Mach. Intel. **28**, 1393 (2006)
10. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002)
11. M.E.J. Newman, M. Girvan, Phys. Rev. E **69**, 026113 (2004)
12. A. Clauset, M.E.J. Newman, C. Moore, Phys. Rev. E **70**, 066111 (2004)
13. M.E.J. Newman, Phys. Rev. E **69**, 066133 (2004)
14. M.E.J. Newman, Eur. Phys. J. B **38**, 321 (2004)
15. M.E.J. Newman, Phys. Rev. E **74**, 036104 (2006)
16. M.E.J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006)
17. M.E.J. Newman, E.A. Leicht, Proc. Natl. Acad. Sci. USA **104**, 9564 (2007)
18. H. Zhou, Phys. Rev. E **67**, 041908 (2003)
19. H. Zhou, Phys. Rev. E **67**, 061901 (2003)
20. F. Wu, B.A. Huberman, Eur. Phys. J. B **38**, 331 (2004)
21. J. Duch, A. Arenas, Phys. Rev. E **72**, 027104 (2005)
22. R. Guimera, L.A.N. Amaral, Nature **433**, 895 (2005)
23. G. Palla, I. Derényi, I. Farkas, T. Vicsek, Nature **435**, 814 (2005)

24. L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, J. Stat. Mech. **9**, P09008 (2005)
25. S. Zhang, R.S. Wang, X.S. Zhang, Physica A **374**, 483 (2007)
26. S. Zhang, X.M. Ning, X.S. Zhang, Eur. Phys. J. B **57**, 67 (2007)
27. J.M. Hofman, C.H. Wiggins, Phys. Rev. Lett. **100**, 258701 (2008)
28. M. Rosvall, C.T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008)
29. W. E, T. Li, E. Vanden-Eijnden, Proc. Natl. Acad. Sci. USA **105**, 7907 (2008)
30. T. Li, J. Liu, W. E, Phys. Rev. E **80**, 026106 (2009)
31. A. Lancichinetti, S. Fortunato, J. Kertész, New J. Phys. **11**, 033015 (2009)
32. A. Lancichinetti, S. Fortunato, F. Radicchi, Phys. Rev. E **78**, 46110 (2008)
33. A. Lancichinetti, S. Fortunato, Phys. Rev. E **80**, 16118 (2009)
34. A. Lancichinetti, S. Fortunato, Phys. Rev. E **80**, 56117 (2009)
35. S. Fortunato, Phys. Rep. **486**, 75 (2010)
36. W.H.A. Schilders, H.A. Van der Vorst, J. Rommes, *Model Order Reduction: Theory, Research Aspects and Applications* (Springer, Berlin, Heidelberg, 2008)
37. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001)
38. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. **21**, 1087 (1953)
39. S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science **220**, 671 (1983)
40. X.L. Xie, G. Beni, IEEE Tran. Pattern Anal. Mach. Intel. **13**, 841 (1991)
41. J. Noh, H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004)
42. L. Lovasz, Combinatorics, Paul Erdös is Eighty **2**, 1 (1993)
43. F.R.K. Chung, *Spectral Graph Theory* (American Mathematical Society, Rhode Island, 1997)
44. W.W. Zachary, J. Anthrop. Res. **33**, 452 (1977)
45. D. Lusseau, Proc. R. Soc. B: Biol. Sci. **270**, 186 (2003)
46. D. Lusseau, K. Schneider, O.J Boisseau, P. Haase, E. Slooten, S.M. Dawson, Behav. Ecol. Sociobiol. **54**, 396 (2003)