Physica A 390 (2011) 3579–3591

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A validity index approach for network partitions

Jian Liu, Tiejun Li [*]

LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, PR China

ARTICLE INFO

Article history:
Received 16 September 2010
Received in revised form 8 February 2011
Available online 1 May 2011

Keywords:
Model selection
Network reduction
Markov chain

ABSTRACT

The validity index has been used to evaluate the fitness of partitions produced by clustering algorithms for points in Euclidean space. In this paper, we propose a new validity index for network partitions, which can provide a measure of goodness for the community structure of networks. It is defined as a product of two factors, and involves the compactness and separation for each partition. The simulated annealing strategy is used to minimize such a validity index function in coordination with our previous k-means algorithm based on the optimal reduction of a random walker Markovian dynamics on the network. It is demonstrated that the algorithm can efficiently find the community structure during the cooling process. The number of communities can be automatically determined without any prior knowledge of the community structure. Moreover, the algorithm is successfully applied to three real-world networks.

1. Introduction

In recent years we have seen an explosive growth of interest and activity concerning the structure and dynamics of complex networks [1–4]. 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 [5–7]. 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 [8–10]. 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 networks into a small number of communities [11–23], which are constructed from different viewpoints. The state of the art of community detection methods was recently summarized in [24]. In a broader aspect, it is also closely related to the model reduction theory of differential equations [25].

In a previous paper [22], the authors proposed an approach to partition the networks based on the optimal approximation of a transition matrix. The basic idea is to associate the network with a random walker Markovian dynamics [26], and then introduce a metric (Hilbert–Schmidt metric for the forward operator) on the space of Markov chains, and optimally reduce the chain under this metric. The final minimization problem is solved by an analogy to the traditional k-means algorithm in clustering analysis [27]. This approach also bears some similarity to the MNCut algorithms used in image segmentation [8,9] and the diffusion maps used in data mining [10].

In traditional clustering literature, the family of standard k-means algorithms is based on the optimization of a specified objective function with known number of clusters [27]. However, one is sometimes required to determine the number of

[*] Corresponding author.
E-mail addresses: dugujian@pku.edu.cn (J. Liu), tieli@pku.edu.cn, tieli@math.pku.edu.cn (T. Li).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.04.015

communities of the optimal network partition, and this encounters the difficulty that the objective function in k-means usually decreases as the number of communities increases. This motivates the idea of constructing a function of validity index [28–42] to evaluate the quality of clustering results. The optimal number of clusters can be determined by selecting the minimum (or maximum by different definitions) value of the index. Using the same idea, we construct a new validity index, which involves the compactness and separation for each partition to measure the goodness of the network community structure. Then a simulated annealing strategy [43,44] is utilized to obtain the minimum of this function. This kind of simulated annealing in coordination with our previous k-means iteration has a high degree of efficiency and accuracy since the process of iteration can accelerate the tendency of minimizing the validity index. This approach can not only identify the community structure efficiently, but also determine the optimal number of communities automatically without any prior knowledge about the community structure. The proposed validity index is competitive with the modularity function for network community structure in the literature [14–17]. Furthermore, it has more predictive power than some other ways of doing network partition.

We construct our algorithm – *S*imulated *A*nnealing to minimize *V*alidity *I*ndex (SAVI) associating with a k-means iteration – for network partitioning. The algorithm is tested on several artificial networks, including an ad hoc network, the sample network generated from a Gaussian mixture model, and the Lancichinetti–Fortunato–Radicchi (LFR)-benchmarks [45]. The numerical results suggest that our algorithm is efficiently implemented with reasonable computational effort and leads to accurate partitioning results. Moreover, successful application to three real-world networks, namely a karate club network, a dolphin network, and an American football team network, confirms the effectiveness of the present algorithm. We would also like to remark that though the validity index is not a new concept, application to the reduction of Markov chains and the detection of the community structure of complex networks is novel to the best knowledge of the authors. It is also a natural extension of the former k-means algorithm in the context of Markov chain aggregation.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the framework of network partition based on optimal prediction theory. In Section 3, we describe our validity index for network partition in detail. The algorithm (SAVI) and the corresponding strategies are proposed in Section 4. In Section 5, we apply SAVI to the representative examples mentioned above. The partitioning results are typically compared. Finally, we present our conclusions in Section 6.

2. Network partition based on optimal prediction

In [22], a new strategy for reducing the random walker Markovian dynamics based on the optimal approximation of a transition matrix is proposed. Let $G(S, E)$ be a network with n nodes and m edges, where S is the node set, $E = \{e(x, y)\}_{x,y\in S}$ is the weight matrix, and $e(x, y)$ is the weight for the edge connecting nodes x and y. We can relate this network to a discrete-time Markov chain with stochastic matrix $P = (p(x, y))$ whose entries are 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 node x [26,46]. This Markov chain has stationary distribution

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

and it satisfies the detailed balance condition $\mu(x)p(x, y) = \mu(y)p(y, x)$.

The basic idea in [22] is to introduce a metric (Hilbert–Schmidt metric for the forward operator associated with P in the Hilbert space $L^2(\mu)$) 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 $\mathbb{S} = \{S_1, \dots, S_N\}$. This matrix can be naturally lifted to the space of stochastic matrices on the original state space S via

$$\tilde{p}(x,y) = \sum_{k,l=1}^{N} \mathbf{1}_{S_k}(x)\hat{p}_{kl}\mu_l(y), \tag{4}$$

where $\mathbf{1}_{S_k}(x) = 1$ if $x \in S_k$ and $\mathbf{1}_{S_k}(x) = 0$ otherwise, and

$$\mu_k(x) = \frac{\mu(x)\mathbf{1}_{S_k}(x)}{\hat{\mu}_k}, \qquad \hat{\mu}_k = \sum_{z\in S_k} \mu(z). \tag{5}$$

Based upon this formulation, we can find the optimal $\hat{p}_{kl}$ for any fixed partition. With this optimal form $\hat{p}_{kl}$, we further search for the best partition $\{S_1, \ldots, S_N\}$ with the given number of communities N by minimizing the optimal prediction error

$$\min_{\{S_1,\ldots,S_N\},\hat{p}_{kl}} J = \|\tilde{p} - p\|_\mu^2 = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} \Big[\tilde{p}(x,y) - p(x,y)\Big]^2$$
$$= \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} p^2(x,y) - \sum_{k,l=1}^{N} \frac{\hat{\mu}_k}{\hat{\mu}_l} \hat{p}_{kl}^2. \tag{6}$$

A direct calculation shows that when the partition is known the minimizer of Eq. (6) is unique, and can be given as

$$\hat{p}_{kl} = \sum_{x\in S_k, y\in S_l} \mu_k(x) p(x,y). \tag{7}$$

It can be checked that Eq. (7) is a stochastic matrix and that $\hat{\mu}$ in Eq. (5) is an equilibrium distribution for the Markov chain on $\mathbb{S}$ with transition matrix given in Eq. (7). Furthermore, it is easy to see that Eq. (7) satisfies the detailed balance condition with respect to $\hat{\mu}$. An analogy to the k-means algorithm can be applied to minimize the combinatorial optimization problem (6). Given an initial partition $\{S_k^{(0)}\}_{k=1}^N$, for the t-th step, use

$$S_k^{(t+1)} = \Big\{x : k = \arg\min_l D(x, S_l^{(t)})\Big\} \tag{8}$$

to update the new state, where

$$D(x, S_k) = \sum_{l=1}^{N} \sum_{y\in S_l} \mu(x)\mu(y) \left(\frac{p(x,y)}{\mu(y)} - \frac{\hat{p}_{kl}}{\hat{\mu}_l}\right)^2. \tag{9}$$

This is the theoretical basis for constructing the k-means algorithm for the community structure of complex networks in [22]. The framework is further extended successfully to fuzzy formulations [23,47].

3. The validity index criterion

The validity index is a quantity that can measure how well the clustering results reflect the structure of the data set S. The most important indicator of the structure is the number of clusters; most basic clustering algorithms assume that this is a user-defined parameter. However, the number of clusters is a parameter related to the complexity of the data structure. In other words, the clustering algorithm is run with different initial values for the number of clusters and the results are compared in order to determine the appropriate number of clusters. For this purpose, various validity indices have been developed in the literature. Some existing definitions both for hard and fuzzy clustering are briefly reviewed as follows, but this is not designed to be a complete list. In our proposed new validity index, we mainly mimic Xie and Beni's definition [35] because of its simple form and its effectiveness, which is justified from the numerical results in Section 5. The Xie–Beni index is more accurate compared with some other indices proposed before, and many other validity indices constructed subsequently have also been derived from it [39–41].

3.1. Validity indices for hard clustering

3.1.1. The Dunn index

A well-established crisp cluster validity index is the separation index V_D [28], which identifies compact and separate clusters, and is defined by

$$V_D = \min_{1\le k\le N} \left\{ \min_{k+1\le l\le N-1} \left\{ \frac{\mathrm{dis}(S_k, S_l)}{\max\limits_{1\le m\le N} \{\mathrm{dia}(S_m)\}} \right\} \right\}, \tag{10}$$

where

$$\mathrm{dia}(S_k) = \max_{x,y\in S_k} \|x - y\|, \tag{11}$$

$$\mathrm{dis}(S_k, S_l) = \min_{x\in S_k, y\in S_l} \|x - y\|. \tag{12}$$

Here, $\|\cdot\|$ is any metric induced by an inner product on $\mathbb{R}^n$. The compact and separate clustering of S is to be found by solving $\max_{2\le N\le n} \big\{\max_{\{S_1,\ldots,S_N\}} V_D\big\}$, where $\{S_1, \ldots, S_N\}$ denotes the optimal partition at fixed N. It is proved [28] that a crisp partition of S contains N compact and separate clusters if $V_D > 1$. Furthermore, there is at most one partition of S if $V_D > 1$. The main drawback with direct implementation of this validity index is in computations, since calculating V_D becomes computationally very expensive as N and n increase.

3.1.2. The Davies–Bouldin index

Another validity index which also measures compact and separate clusters was introduced by Davies and Bouldin [29]. This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within the k-th cluster is computed as

$$SC_k = \frac{1}{|S_k|} \sum_{x \in S_k} \|x - m_k\|, \tag{13}$$

where $|S_k|$ is the number of data points in S_k and m_k is the cluster centroid. The centroid distance is

$$d_{kl} = \|m_k - m_l\|. \tag{14}$$

The Davies–Bouldin index V_{DB} is then defined as

$$V_{\mathrm{DB}} = \frac{1}{N} \sum_{k=1}^{N} \max_{l:l \neq k} \left\{ \frac{SC_k + SC_l}{d_{kl}} \right\}. \tag{15}$$

The objective is to minimize V_{DB} to achieve proper clustering. Its major difference from V_D is that it considers the average case by using the average error of each cluster.

3.2. Validity indices for soft clustering

One popularly used method to make soft clusterings in Euclidean space is the fuzzy c-means algorithm [48,49]. The main idea of the traditional fuzzy c-means algorithm is to minimize the objective function

$$J(\rho, m) = \sum_{k=1}^{N} \sum_{x \in S} \rho_k^b(x) \|x - m_k\|^2, \quad b \geq 1, \tag{16}$$

where x are samples and m_k are centers. $b = 2$ is often chosen in computations. $\rho_k(x)$ denotes the probability of x belonging to cluster k, which satisfies the condition

$$\rho_k(x) \geq 0, \qquad \sum_{k=1}^{N} \rho_k(x) = 1, \quad x \in S. \tag{17}$$

The proposed validity indices for soft clustering problems in the literature are of many forms, such as the partition entropy, Fukuyama–Sugeno index, and the fuzzy hypervolume validity [33,34,37–42]. But this is far from a complete list since different constructions are proposed from different motivations. We only list two of them in what follows because of their simple geometric intuitions.

3.2.1. The Fukuyama–Sugeno index

The validity function proposed by Fukuyama and Sugeno [34] was defined by

$$\begin{aligned} V_{\mathrm{FS}} &= \sum_{x \in S} \sum_{k=1}^{N} \rho_k^2(x) \|x - m_k\|^2 - \sum_{x \in S} \sum_{k=1}^{N} \rho_k^2(x) \|m_k - \bar{m}\|^2 \\ &= J(\rho, m) + K(\rho, m), \end{aligned} \tag{18}$$

where $\bar{m} = \sum_{k=1}^{N} m_k / N$. Here, $J(\rho, m)$ is the objective function of Fuzzy C-Means (FCM) algorithm with $b = 2$, which measures the compactness, and $K(\rho, m)$, which measures the separation. So the goal is to find the fuzzy partition with the smallest V_{FS}.

3.2.2. The Xie–Beni index

Another famous validity index, called the Xie–Beni index [35], can be explicitly written as

$$V_{\mathrm{XB}} = \frac{\sum_{x \in S} \sum_{k=1}^{N} \rho_k^2(x) \|x - m_k\|^2}{n \min_{k \neq l} \|m_k - m_l\|^2} = \frac{J(\rho, m)}{nK(m)}. \tag{19}$$

More importantly, minimizing V_{XB} corresponds to minimizing $J(\rho, m)$, which is the goal of FCM with $b = 2$. The additional factor is $K(m)$, which is the separation measurement. The more separate the clusters, the larger $K(m)$ and the smaller V_{XB}. More information about it may be found in [36].

3.3. The new validity index for network partition

We take the idea of considering both compactness and separation in our formulation, and construct a validity index for network partition as follows:

$$V_{\text{net}} = J(\hat{p}) \cdot K(\hat{p}), \tag{20}$$

where $J(\hat{p})$ is the objective function in Eq. (6) which reflects compactness, and the term

$$K(\hat{p}) = \frac{N - \sum\limits_{k=1}^{N} \hat{p}_{kk}}{\sum\limits_{k=1}^{N} \hat{p}_{kk}} \cdot \frac{1}{N-1} = \frac{\sum\limits_{k \neq l} \hat{p}_{kl}}{\sum\limits_{k=1}^{N} \hat{p}_{kk}} \cdot \frac{1}{N-1} = \frac{\frac{1}{N(N-1)} \sum\limits_{k \neq l} \hat{p}_{kl}}{\frac{1}{N} \sum\limits_{k=1}^{N} \hat{p}_{kk}} \tag{21}$$

plays the role of separation such as $K(m)$ in Eq. (19). Here, $\frac{1}{N(N-1)} \sum_{k \neq l} \hat{p}_{kl}$ represents the average of the probability with which one community transits to another and $\frac{1}{N} \sum_{k=1}^{N} \hat{p}_{kk}$ represents the average of the probability with which one community stays by itself. An ideal partition requires a more stable state in space $\mathbb{S} = \{S_1, \dots, S_N\}$, which has smaller $\frac{1}{N(N-1)} \sum_{k \neq l} \hat{p}_{kl}$ and larger $\frac{1}{N} \sum_{k=1}^{N} \hat{p}_{kk}$. Thus, the optimal partition can be found by solving

$$\min_{N} \left\{ \min_{\{S_1,\dots,S_N\}} V_{\text{net}} \right\}. \tag{22}$$

According to Eq. (6), we obtain

$$V_{\text{net}} = \frac{1}{N-1} \left[\sum_{x,y \in S} \frac{\mu(x)}{\mu(y)} p^2(x,y) - \sum_{k,l=1}^{N} \frac{\hat{\mu}_k}{\hat{\mu}_l} \hat{p}_{kl}^2 \right] \cdot \frac{N - \sum\limits_{k=1}^{N} \hat{p}_{kk}}{\sum\limits_{k=1}^{N} \hat{p}_{kk}} \tag{23}$$

as a new validity index for network partition along the line of [22].

As a model selection framework, one will usually encounter tuning parameters, which control the competition between the compactness J and separation K in our setup. This is easy to include in the current case. We define

$$V_{\text{net}}^{\lambda} = J(\hat{p}) \cdot K(\hat{p})^{\lambda}, \quad \lambda \in \mathbb{R}, \tag{24}$$

where λ is the regularization parameter. When $\lambda = 0$, V_{net}^{λ} degenerates to J. In all of the following numerical tests, we will only consider the case $\lambda = 1$, which already gives satisfying results.

4. The algorithm

The first simulated annealing algorithm was motivated by simulating the physical process of annealing solids [43]. The process can be described as follows. First, a solid crystal is heated to a high temperature and then cooled slowly so that the system at any time is approximately in thermodynamic equilibrium. At equilibrium, there may be many configurations, with each one corresponding to a specific energy level. The chance of accepting a change from the current configuration to a new configuration is related to the difference in energy between the two states. The simulated annealing strategy is widely used in optimization problems [44].

Let $E = V_{\text{net}}$. $E^{(t)}$ and $E^{(t+1)}$ represent the current energy and new energy, respectively. $E^{(t+1)}$ is always accepted if it satisfies $E^{(t+1)} < E^{(t)}$, but if $E^{(t+1)} > E^{(t)}$ the new energy level is only accepted with a probability as specified by $\exp\left(-\frac{1}{T}\Delta E^{(t)}\right)$, where $\Delta E^{(t)} = E^{(t+1)} - E^{(t)}$ is the difference of energy and T is the current temperature. Higher energy states are possibly accepted, which allows avoiding being trapped at local minima. The temperature is then decreased gradually and the annealing process is repeated until no more improvement is reached or any termination criterion is met.

At a given temperature, the new state $\{S_k^{(t+1)}\}_{k=1}^{N}$ is accepted with a probability $\exp\left(-\frac{1}{T}\Delta E^{(t)}\right)$, where the energy is used to evaluate a partition. The initial state is generated by the k-means algorithm in [22] with randomly initialized N communities, where N is an integer within the range $[N_{\min}, N_{\max}]$; $N_{\min} = 2$, $N_{\max} = n/3$ is chosen in our computation. The initial temperature T is set to a high temperature, $T_{\max}$. Then the next proposal state is produced by applying the k-means algorithm to the initial state generated by our two proposal operations below. The new state is kept if the acceptance requirement is satisfied. This process will be repeated R times at the given temperature. A cooling rate $0 < \alpha < 1$ is set to decrease the current temperature until the lower bound $T_{\min}$ is reached. The whole procedure of simulated annealing to minimize the validity index (SAVI) associating with the k-means iteration algorithm is summarized as follows.

Algorithm 1 (*SAVI*)**.** Simulated annealing algorithm to minimize the validity index.

(1) Set the parameters $T_{\max}$, $T_{\min}$, α, and R. Choose N randomly within the range $[N_{\min}, N_{\max}]$, and initialize the partition $\{S_k^{(0)}\}_{k=1}^{N}$ randomly. Set the current temperature $T = T_{\max}$.

(2) Compute the corresponding $\hat{p}_{kl}^{(0)}$ and $\{S_k^{(0)}\}_{k=1}^N$ with the k-means algorithm in [22] according to Eqs. (7) and (8). Calculate the energy E^* using the definition of Eq. (23).

(3) For $t = 0, 1, \ldots, R$, do the following iterations.

  (3.1) Generate a new partition $\{S_k^{(i)}\}_{k=1}^{N'}$ as an initial partition according to the proposal below, and set $N = N'$.

  (3.2) Update the coarse-grained transition probability $\hat{p}_{kl}^{(n)}$ and the partition $\{S_k^{(t+1)}\}_{k=1}^N$ with the k-means algorithm in [22] according to Eqs. (7) and (8). Update the new energy $E^{(t+1)}$ according to Eq. (23).

  (3.3) Accept the new partition with the standard Metropolis criterion, i.e. accept with probability $\min\left\{1, \exp\left(-\frac{1}{T}\Delta E^{(t)}\right)\right\}$. Set $t = t + 1$.

  (3.4) Update the optimal state. If $E^{(t)} < E^*$, set $E^* = E^{(t)}$, and record the current partition.

(4) Decrease the temperature to $T = \alpha \cdot T$. If $T < T_{\min}$, go to Step (5); otherwise, repeat Step (3).

(5) Output the optimal partition $\{S_k\}_{k=1}^N$ and the minimum energy E^*.

Our proposal for the process of generating a set of new partitions in Step (3.1) is comprised of two operations, which are deleting a current community and splitting a current community. At each k-means iteration, one of the two operations can be randomly chosen, and the community strength

$$M_k = \hat{p}_{kk}, \quad k = 1, \ldots, N \tag{25}$$

is used to select a community, which reflects the possibility of the k-th community remaining by itself and not tending to transit to the others. Obviously, the clustering tendency of a community is stronger if its strength is larger. The two operations are described below.

(i) Deleting a community. The community with the minimum community strength M_d is identified, where $d = \arg\min_k M_k$. It should be deleted from the current partition and merged to community S_k, where $k = \arg\max_m \hat{p}_{dm}$.

(ii) Splitting a community. The community with the maximum community strength M_s is chosen and randomly split to two new communities with equal size. (If the number of nodes m is odd, the sizes of the two subcommunities will be $(m+1)/2$ and $(m-1)/2$, respectively.)

Note that we can obtain the global minima of (23) by searching over all possible N with k-means. But this will cost too much since for each fixed N the k-means algorithm should be repeated $O(10^2)$ times to get a trustable good partition. However, the annealing procedure simulated above can avoid this repetition by searching for the optimal community number one by one. The numerical performance of the whole algorithm on the sample networks is very efficient and successful.

5. Numerical examples

In this section, we extensively test our algorithm on some artificial networks with a known community structure, including an ad hoc network with 128 nodes, the sample network generated from a Gaussian mixture model, and the LFR benchmarks. Then the algorithm is applied to some real-world networks: the social interactions between members of a karate club, the relationships of bottlenose dolphins living in Doubtful Sound, New Zealand, and the conference connection network of US college football teams.

5.1. *Tests on artificial networks*

5.1.1. *Ad hoc networks with 128 nodes*

In this subsection, we apply our method to an ad hoc network with 128 nodes. The ad hoc network is a typical benchmark problem considered in many papers [13,14,18,19,22,23]. It has a known community structure and is constructed as follows. Suppose that we choose $n = 128$ nodes, split into four communities containing 32 nodes each. Assume that pairs of nodes belonging to the same communities are linked with probability p_{in}, and that pairs belonging to different communities are linked with probability p_{out}. These values are chosen so that the average node degree, d, is fixed at $d = 16$. In other words, p_{in} and p_{out} are related as

$$31p_{\text{in}} + 96p_{\text{out}} = 16. \tag{26}$$

Here, we naturally choose the node group $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. We change $z_{\text{out}} = 96p_{\text{out}}$ from 0.5 to 8 and investigate the fraction of nodes which are correctly classified. By setting $T_{\max} = 3, T_{\min} = 10^{-2}, \alpha = 0.9$ and $R = 20$, we create clustering using SAVI. The fraction of correctly identified nodes is shown in Fig. 1. Comparing with the two methods described in [14], we can see that SAVI performs noticeably better than the two previous methods, especially for the more diffusive cases when z_{out} is large.

To test on well-clustered network, we take $z_{\text{out}} = 5$. When the annealing strategy is not used, i.e. only the k-means algorithm [22] is applied, the change in the validity index V_{net} and the objective function J with the number of communities N is as shown in Fig. 2(a) and Table 1. We can see the optimal community structure is achieved at $N = 4$; the corresponding validity index is $V_{\text{net}} = 0.9281$. Our algorithm identifies the desired result without knowing the number of communities as a prior parameter.

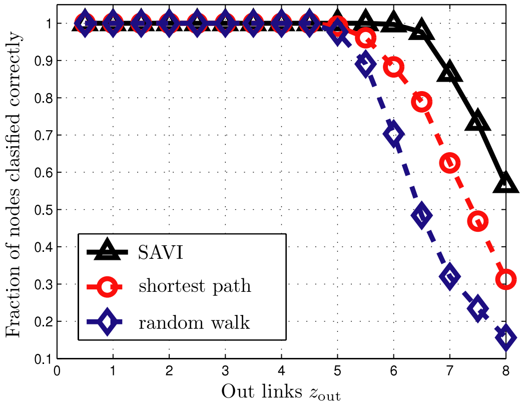

Fig. 1. The fraction of nodes classified correctly by SAVI and the methods used in [14]. SAVI performs better than the shortest path and random walk methods [14] from the figure.

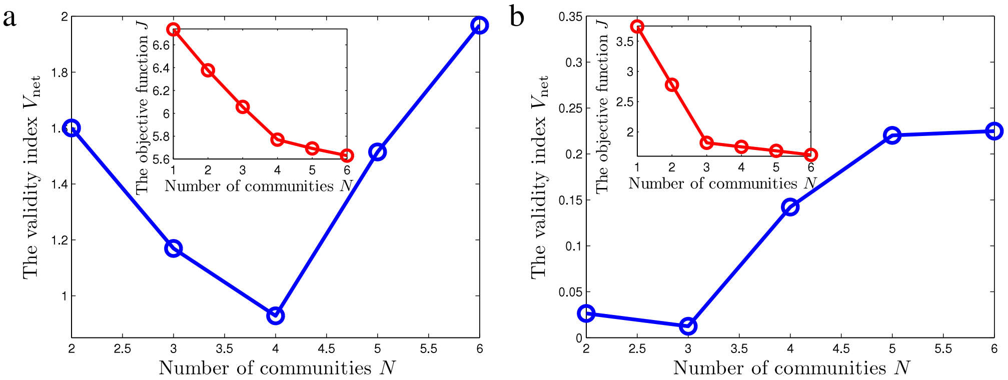

Fig. 2. The change in the validity index V_{net} and the objective function J with the number of communities N for an ad hoc network with $z_{\mathrm{out}} = 5$ and a Gaussian mixture network with 400 nodes.

Table 1
The change in the values of V_{net} and J with the number of communities N for an ad hoc network with $z_{\mathrm{out}} = 5$, a fully connected network with 256 nodes, and a Gaussian mixture network with 400 nodes.

	N	2	3	4	5	6
Ad hoc network	V_{net}	1.6001	1.1694	**0.9281**	1.5143	1.9677
($z_{\mathrm{out}} = 5$)	J	6.3766	6.0560	5.7696	5.6931	5.6293
Gaussian mixture	V_{net}	0.0264	**0.0124**	0.1422	0.2203	0.2249
Network	J	2.7795	1.8218	1.7550	1.6891	1.6223

5.1.2. Sample network generated from the Gaussian mixture model

To further test the validity of the algorithm, we apply it to a sample network generated from a Gaussian mixture model. This model is related to the concept of a random geometric graph proposed by Penrose [50], except that we take a Gaussian mixture here rather than a uniform distribution as in [50].

First, we generate n sample points $\{\boldsymbol{x}_i\}$ in two-dimensional Euclidean space subject to a K-Gaussian mixture distribution

$$\sum_{k=1}^{K} q_k G\left(\boldsymbol{\mu}_k, \Sigma_k\right), \tag{27}$$

where $\{q_k\}$ are mixture proportions satisfying $0 < q_k < 1$, $\sum_{k=1}^{K} q_k = 1$. $\boldsymbol{\mu}_k$ and Σ_k are the mean positions and covariance matrices for each component, respectively. Then we generate the network with a thresholding strategy. That is, if $|\boldsymbol{x}_i - \boldsymbol{x}_j| \leq \mathrm{dist}$, we set an edge between the i-th and j-th nodes; otherwise, they are not connected. With this strategy, the topology of the network is induced by the metric. As a consequence, some properties of the network, say the clustering nature, may be inherited from the case with a metric. This is our basic motivation with this model.

We take $n = 400$ and $K = 3$, then generate the sample points with means and covariance matrices as follows:

$$\boldsymbol{\mu}_1 = (1.0, 4.0)^T, \qquad \boldsymbol{\mu}_2 = (2.5, 5.5)^T, \qquad \boldsymbol{\mu}_3 = (0.5, 6.0)^T, \tag{28}$$

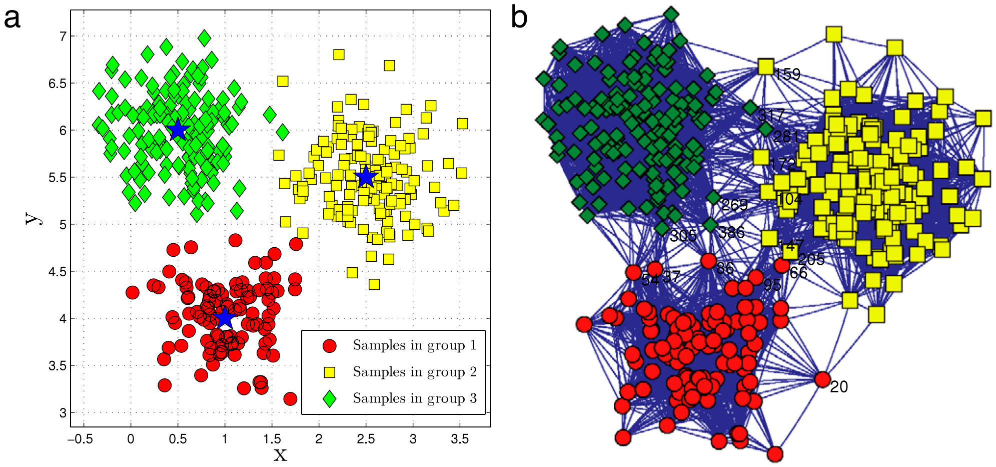

Fig. 3. (a) 400 sample points generated from the given 3-Gaussian mixture distribution. The star symbols represent the centers of each Gaussian component. The circle, square, and diamond-shape symbols represent the positions of the sample points in each component, respectively. (b) The network generated from the sample points in (a) with the parameter dist = 0.8.

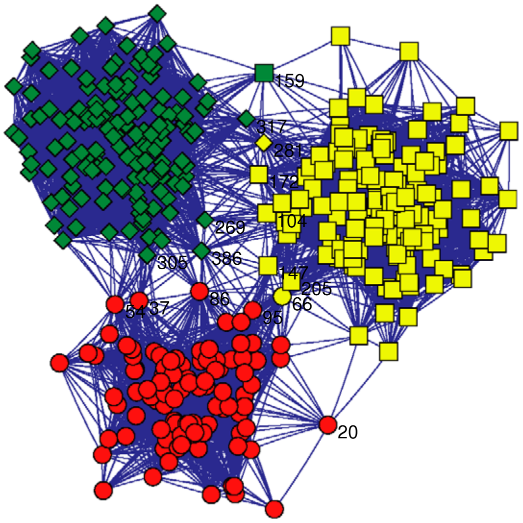

Fig. 4. The community structure for the Gaussian mixture network with 400 nodes obtained by our method. Only nodes 66, 159, and 281 do not coincide with the initial groups of samples generated in Euclidean space.

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{29}$$

Here we pick nodes 1:100 in group 1, nodes 101:250 in group 2 and nodes 251:400 in group 3, for simplicity. With this choice, we have $q_1 = 100/400$, $q_2 = q_3 = 150/400$. We take dist $= 0.8$ in this example. The sample points are shown in Fig. 3(a) and the corresponding network is shown in Fig. 3(b). The change in computed validity index value V_{net} with the number of communities N is shown in Fig. 2(c) and Table 1. We can observe that the optimal community structure is achieved at $N = 3$; the corresponding validity index is $V_{\text{net}} = 0.0124$. By applying SAVI with $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$, and $R = 20$, we also obtain $N = 3$ and $V_{\text{net}} = 0.0124$. The partition result is shown in Fig. 4. Only nodes 66, 159, and 281 do not coincide with the initial groups of samples generated in Euclidean space. Our algorithm achieves a reasonable partitioning result that confirms the intuition from the network topology visualization.

5.1.3. The LFR benchmarks

The LFR benchmark [45,51] 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

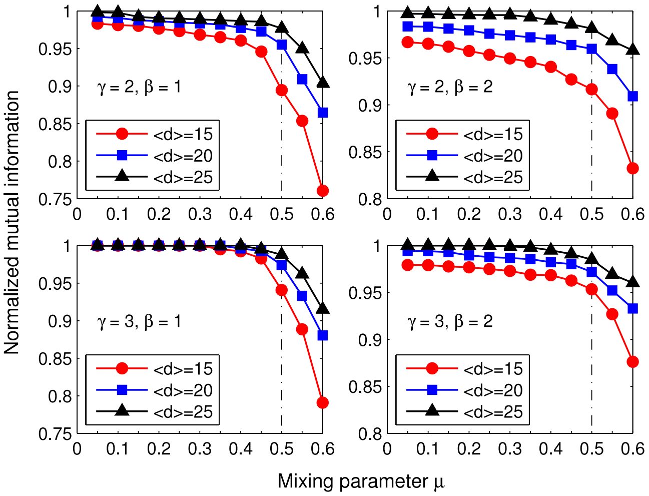

Fig. 5. Test of our algorithm on the LFR benchmark problem [45,51]. The number of nodes $n = 500$. The results clearly depend on all parameters of the benchmark, from the exponents γ and β to the average degree $\langle d \rangle$. The threshold $\mu_c = 0.5$, shown by the dashed vertical line in the 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. Each point corresponds to an average over 20 graph realizations. The overall results show that our algorithm gives very good accuracy for the detection of the communities. For the normalized mutual information, our results are all above 0.9 when $\mu \le \mu_c$, and they are also very competitive for the more diffusive cases.

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. To compare the built-in modular structure with the one delivered by the algorithm, we adopt the normalized mutual information, which has proved to be reliable [19,45,51]. It is based on defining a confusion matrix M, in which the rows correspond to the real communities, and the columns correspond to the found communities. The component 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 as N_r and the number of found communities is denoted as N_f, the sum over row k of matrix M_{kl} is denoted as M_k, and the sum over column l is denoted as M_l. The measure of similarity between the partitions, based on information theory, is given as

$$I(\mathbb{S}_r, \mathbb{S}_f) = \frac{-2\sum_{k=1}^{N_r}\sum_{l=1}^{N_f} M_{kl}\log\left(\frac{nM_{kl}}{M_k M_l}\right)}{\sum_{k=1}^{N_r} M_k \log\left(\frac{M_k}{n}\right) + \sum_{l=1}^{N_f} M_l \log\left(\frac{M_l}{n}\right)}. \tag{30}$$

In Fig. 5, we show the results when we apply our algorithm to the benchmark problem with $n = 500$. The parameters are set as $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$, and $R = 20$. The four panels shown in the figure correspond to the results with the four pairs of exponents $(\gamma, \beta) = (2, 1), (2, 2), (3, 1)$, and $(3, 2)$, respectively. We have chosen combinations of the extremes of the exponent 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 our method is better when the average degree $\langle d \rangle$ is larger, whereas it gets worse when the mixing parameter becomes 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 Fig. 6, we compare our algorithm with the Infomap algorithm [21] for the case of $\langle d \rangle = 20$, and the other parameters are chosen as $(\gamma, \beta) = (2, 1), (2, 2)$. Our result shows that it is very competitive with the Infomap algorithm, especially for the more diffusive cases when the mixing parameter μ is large. These results also support the effectiveness of our algorithm.

5.2. Application to real-world networks

5.2.1. 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 [52]. Soon after, a dispute arose between the club's administrator and the main teacher, and the club split into two smaller clubs. It has been used widely to test the algorithms for finding communities in networks [13–18,22,23].

The change in validity index V_{net} and the objective function J with the number of communities N is shown in Fig. 7(a) and Table 2 when only the k-means algorithm is applied. The community structures with $N = 2$ and $N = 3$ are shown in Fig. 8. We can find that the optimal community structure is achieved at $N = 3$; the corresponding validity index is $V_{\text{net}} = 0.4711$.

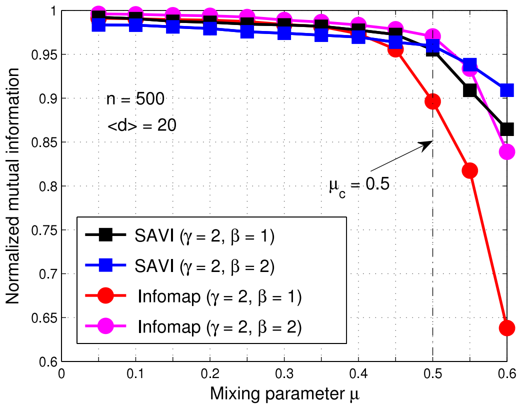

Fig. 6. Test of our algorithm compared with the Infomap algorithm [21] on the LFR benchmark [45,51]. The number of nodes $n = 500$ and the average degree $\langle d \rangle = 20$. Our method shows that our algorithm is very competitive with the Infomap algorithm. When μ is small, both algorithms give very good accuracy, which is close to 1 in terms of the normalized mutual information. For the more diffusive cases ($\mu > \mu_c = 0.5$), our algorithm performs better than Infomap.

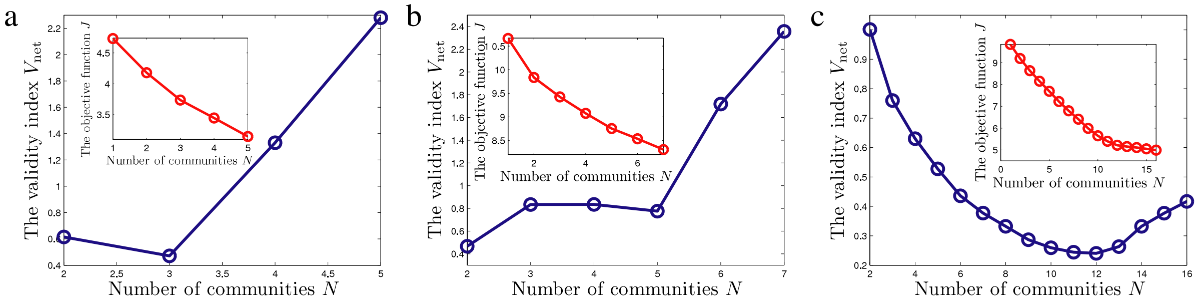

Fig. 7. The change in validity index V_{net} and the objective function J with the number of communities N for the karate club network, the dolphin network, and the football team network. The global minimum shows the optimal number of communities obtained by SAVI.

Table 2
The changes in values of V_{net} and J with the number of communities N for the three real-world networks: the karate club network, the dolphin network, and the football team network.

The karate club network			The dolphin network			The football team network		
N	V_{net}	J	N	V_{net}	J	N	V_{net}	J
2	0.6147	4.1798	2	**0.4667**	9.8349	10	0.2594	5.6511
3	**0.4711**	3.7372	3	0.8344	9.4243	11	0.2444	5.3985
4	1.3308	3.4463	4	0.8349	9.0751	12	**0.2403**	5.2169
5	2.2806	3.1472	5	0.7757	8.7538	13	0.2634	5.1557

Note that this result is different from the original partition in Zachary's observation. But from the network topology and the final partition (Fig. 8(b)), it is also reasonable. We remark here that four communities are obtained in [14] with maximizing modularity. If we use SAVI with $T_{max} = 3$, $T_{min} = 10^{-2}$, $\alpha = 0.9$, and $R = 50$, we obtain the same partition when $N = 3$. This phenomenon is closely related to the diffusive nature of this karate club network.

5.2.2. Dolphin network

The dolphin network is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand [53,54]. The network was compiled from the studies of the dolphins, with ties between dolphin pairs being established by observation of statistically significant frequent association [14–16].

When only the k-means algorithm [22] is applied, the change in validity index V_{net} and the objective function J with the number of communities N is as shown in Fig. 7(b) and Table 2. We can see the global optimal community structure is achieved at $N = 2$. When SAVI is used with $T_{max} = 3$, $T_{min} = 10^{-2}$, $\alpha = 0.9$, and $R = 20$, we obtain $N = 2$, and the corresponding $V_{net} = 0.4667$. The partitioning result is shown in Fig. 9. According to the result from using SAVI, the network seems to split into two large communities, shown by the yellow part and the red one, which corresponds to a known division of the dolphin community [53,54]. This illustrates that the validity index can reflect the internal community character effectively.

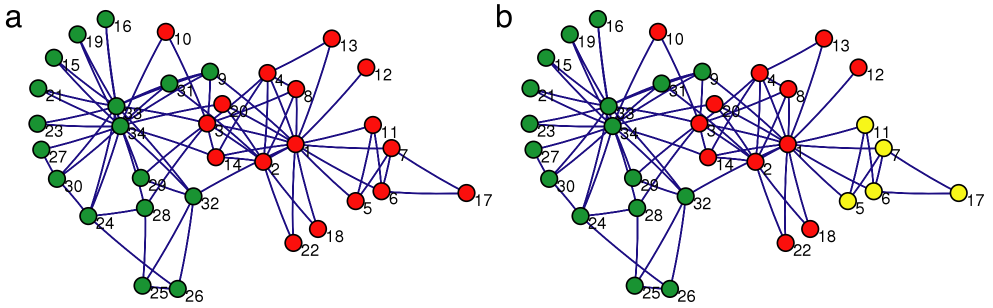

Fig. 8. The community structure of the karate club network obtained by k-means in [22]. (a) The result with $N = 2$. (b) The result with $N = 3$; our method (SAVI) produces the same partition.

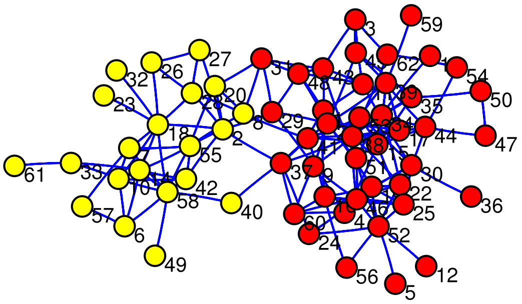

Fig. 9. The nodes shown in the figure with the yellow and red colors correspond to the obtained partition. The community structure of the dolphin network obtained by our method coincides with a known division [54]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.2.3. The football team network

The last network we investigated is a college football network, which represents the game schedule of the 2000 season of Division I of the US college football league [13,20]. The nodes in the network represent 115 teams and the edges represent regular season games between the two teams they connect. The teams are divided into conferences containing around 8–12 teams each. Games are more frequent between members of the same conference than between members of different conferences. Such known community structure makes this network interesting to investigate.

When only the k-means algorithm is applied, the change in validity index V_{net} and the objective function J with the number of communities N is as shown in Fig. 7(c) and Table 2. The global optimal community structure is achieved at $N = 12$. Then SAVI is used with $T_{\max} = 3$, $T_{\min} = 10^{-3}$, $\alpha = 0.9$, and $R = 50$; we obtain $N = 12$, and the corresponding $V_{\text{net}} = 0.2403$. The partitioning result is shown in Fig. 10. According to the result of our method, we identify the community structure with a high degree of accuracy. Almost all of the football teams are correctly clustered with the others in their conference. The teams in the Independents conference, which are denoted as a green-edge box, seem not to belong to any community, but they tend to be clustered with the conference which they are most closely associated with. The Sunbelt conference (shaded diamond) is split into two communities, and each is clustered with a team which is less connected in the Western Athletic conference (shaded triangle). Only one member in Conference USA (shaded black-edge box), Texas Christian, is grouped with most of the teams in the Western Athletic conference (shaded triangle). All the other communities (shaded colored ellipse) coincide with the known structure, which indicates that our algorithm performs remarkably well.

6. Conclusions

In this paper, we have proposed a new validity index function to evaluate the goodness of community structure in networks. The proposed algorithm—simulated annealing to minimize the validity index (SAVI), in coordination with a k-means iteration—is constructed and successfully applied in several representative networks. The experiments on artificial networks with a known structure show very satisfactory results, namely that our algorithm can identify the communities in networks with a high degree of efficiency and accuracy. It can identify the community structure with random initial partition during the cooling process. The optimal number of communities can be automatically determined without any prior knowledge about the community structure. The proposed validity index is competitive with the modularity function for network community structure in the literature [14–17] and has more predictive power than some other ways of doing network partition. Moreover, successful applications to three real-world networks, namely a karate club network, a dolphin network and an American football team network, confirm the effectiveness of the present algorithm.

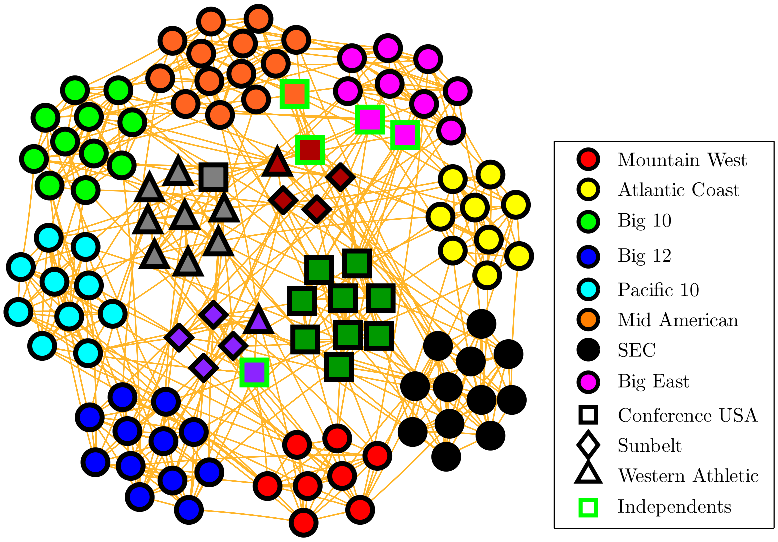

Fig. 10. The community structure of the football team network obtain by our algorithm. Nodes in the network represent teams and edges represent games between teams. The 12 real conferences are represented by different symbols listed in the right box for reference. Our algorithm identifies nearly all the communities in the network which are represented by different colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgments

This work is supported by the National 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, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97.
[2] M. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167–256.
[3] M. Newman, A. Barabási, D. Watts, The Structure and Dynamics of Networks, Princeton University Press, Princeton, 2005.
[4] N.R. Council, Network Science, Natl. Acad. Press, Washington, DC, 2005.
[5] A. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311 (2002) 590–614.
[6] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, A. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (2002) 1551–1555.
[7] G. Flake, S. Lawrence, C. Giles, F. Coetzee, Self-organization and identification of web communities, IEEE Comput. 35 (2002) 66–71.
[8] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 888–905.
[9] M. Meilă, J. Shi, A random walks view of spectral segmentation, in: Proc. 8th International Workshop on Artificial Intelligence and Statistics, Kaufmann, San Francisco, 2001, pp. 92–97.
[10] S. Lafon, A. Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1393–1403.
[11] H. Zhou, Network landscape from a Brownian particle's perspective, Phys. Rev. E 67 (4) (2003) 041908.
[12] H. Zhou, Distance, dissimilarity index, and network community structure, Phys. Rev. E 67 (6) (2003) 061901.
[13] M. Girvan, M. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.
[14] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[15] M. Newman, Detecting community structure in networks, Eur. Phys. J. B 38 (2) (2004) 321–330.
[16] M. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[17] M. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103 (23) (2006) 8577–8582.
[18] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[19] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech. (2005) P09008.
[20] J. Hofman, C. Wiggins, A Bayesian approach to network modularity, Phys. Rev. Lett. 100 (2008) 258701.
[21] M. Rosvall, C. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. USA 105 (4) (2008) 1118–1123.
[22] W. E, T. Li, E. Vanden-Eijnden, Optimal partition and effective dynamics of complex networks, Proc. Natl. Acad. Sci. USA 105 (23) (2008) 7907–7912.
[23] T. Li, J. Liu, W. E, Probabilistic framework for network partition, Phys. Rev. E 80 (2009) 026106.
[24] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[25] W. Schilders, H. Van der Vorst, J. Rommes, Model Order Reduction: Theory, Research Aspects and Applications, Springer, Berlin, Heidelberg, 2008.
[26] L. Lovasz, Random walks on graphs: a survey, Combinatorics 2 (1993) 1–46. Paul Erdös is Eighty.
[27] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.
[28] J. Dunn, Well separated clusters and optimal fuzzy partitions, Cybern. Syst. 4 (1) (1974) 95–104.
[29] D. Davies, D. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[30] J. Bezdek, Cluster validity with fuzzy sets, Cybern. Syst. 3 (3) (1973) 58–73.
[31] R. Dave, Validating fuzzy partitions obtained through c-shells clustering, Pattern Recognit. Lett. 17 (6) (1996) 613–623.
[32] M. Roubens, Pattern classification problems and fuzzy sets, Fuzzy Sets and Systems 1 (4) (1978) 239–253.
[33] J. Bezdek, Mathematical models for systematics and taxonomy, in: Proc. 8th Int. Conf. on Numerical Taxonomy, 1975, pp. 143–166.
[34] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, in: Proc. 5th Fuzzy Syst. Symp., 1989, pp. 247–250.
[35] X. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Tran. Pattern Anal. Mach. Intell. 13 (8) (1991) 841–847.

[36] N. Pal, J. Bezdek, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Syst. 3 (3) (1995) 370–379.
[37] I. Gath, A. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 773–780.
[38] M. Ramze Rezaee, B. Lelieveldt, J. Reiber, A new cluster validity index for the fuzzy c-mean, Pattern Recognit. Lett. 19 (3–4) (1998) 237–246.
[39] N. Zahid, M. Limouri, A. Essaid, A new cluster-validity for fuzzy clustering, Pattern Recognit. 32 (7) (1999) 1089–1097.
[40] M. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognit. 37 (3) (2004) 487–501.
[41] H. Sun, S. Wang, Q. Jiang, FCM-based model selection algorithms for determining the number of clusters, Pattern Recognit. 37 (10) (2004) 2027–2037.
[42] K. Wu, M. Yang, A cluster validity index for fuzzy clustering, Pattern Recognit. Lett. 26 (9) (2005) 1275–1291.
[43] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087.
[44] S. Kirkpatrick, C. Gelatt, M. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[45] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E 78 (4) (2008) 046110.
[46] F. Chung, Spectral Graph Theory, American Mathematical Society, Rhode Island, 1997.
[47] M. Sarich, C. Schütte, E. Vanden-Eijnden, Optimal fuzzy aggregation of networks, Multiscale Model. Simul. 8 (4) (2010) 1535–1561.
[48] J. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Cybern. Syst. 3 (3) (1973) 32–57.
[49] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[50] M. Penrose, Random Geometric Graphs, Oxford University Press, Oxford, 2003.
[51] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (5) (2009) 056117.
[52] W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473.
[53] D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. Lond. B 270 (2003) 186–188.
[54] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, S. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (4) (2003) 396–405.