Physica A 389 (2010) 2300–2309

Contents lists available at ScienceDirect

Physica A

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

Detecting community structure in complex networks using simulated annealing with *k*-means algorithms

Jian Liu [a,*], Tingzhan Liu [b]

[a] *LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, PR China*

[b] *School of Sciences, Communication University of China, Beijing 100024, PR China*

ARTICLE INFO

Article history:
Received 16 November 2009
Received in revised form 5 January 2010
Available online 17 February 2010

Keywords:
Complex networks
Community structure
Simulated annealing
k-means
Modularity

ABSTRACT

Identifying the community structure in a complex network has been addressed in many different ways. In this paper, the simulated annealing strategy is used to maximize the modularity of a network, associating with a dissimilarity-index-based and with a diffusion-distance-based *k*-means iterative procedure. The proposed algorithms outperform most existing methods in the literature as regards the optimal modularity found. They can not only identify the community structure, but also give the central node of each community during the cooling process. An appropriate number of communities can be efficiently determined without any prior knowledge about the community structure. The computational results for several artificial and real-world networks confirm the capability of the algorithms.

1. Introduction

In recent years we have seen an explosive growth of interest and activity as regards the structure and dynamics of complex networks [1,2]. 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, etc, due to our increased capability of analyzing these models [3,4]. 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 small numbers of communities [5–16], which are constructed from different viewpoints, comparing different proposals in the literature.

In a previous work [10], a dissimilarity index for each pair of nodes is proposed, with which one can measure the extent of proximity between nodes of a network and signify to what extent two nodes would 'like' to be in the same community. The basic idea is to associate the network with the random walker Markovian dynamics [17]. This can motivate us to solve the partitioning problem by analogy with the traditional *k*-means algorithm [18] under this measure. Another work [11] is also along the lines of random walker Markovian dynamics, but then introduces the diffusion distance on the space of nodes and identifies the geometric centroid in the same framework. This proximity reflects the connectivity of nodes in a diffusion process. The final minimization problem under this distance can also be solved by a *k*-means algorithm [11].

In traditional clustering literature, the family of standard *k*-means algorithms is based on the optimization of a specified objective function with the known number of clusters [18]. However, people are sometimes required to determine the number of communities of the optimal network partition and encounter the difficulty that the objective function in *k*-means

* Corresponding author. Tel.: +86 010 58570277.
E-mail addresses: dugujian@pku.edu.cn (J. Liu), tzliu@jlu.edu.cn (T. Liu).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.01.042

approaches usually decreases as the number of communities increases. To overcome this weakness, we choose a widely used concept of modularity [6–9] as a valid measure for network partitioning, which has larger values indicating stronger community structure. Then simulated annealing [19,20] can be used to search for the maximal value of the modularity. The cooling process is operated with two kinds of k-means iterations based on the above measures on networks. Such simulated annealing with k-means algorithms is first proposed here and quite different with the previous work [13], since the process of iteration accelerates the tendency of maximizing the modularity function. The algorithms outperform the existing methods in the literature [5–10,12,14] as regards the optimal modularity found. Another advantage is that this category of method can not only identify the community structure, including the number of communities, but also give the central node of each community. The center of a community can convey the information of how important a status it has among the members of the same group, since people are sometimes interested in the characterization of the communication in small groups and assume a relation between structural centrality and influence in group processes [21,22].

We constructed our algorithms — simulated annealing with dissimilarity-index-based k-means (SADI) and simulated annealing with diffusion-distance-based k-means (SADD) — for network partition. The algorithms are tested on two artificial networks, including the ad hoc network and the sample network generated from a Gaussian mixture model. Both methods are efficiently implemented with reasonable computational effort and lead to accurate partitioning results. Moreover, they are successfully applied to several real-world networks, including the karate club network, the dolphins network, the political books network, the network of characters in the novel *Les Misérables* and the American football team network.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the two measures for proximity of nodes in networks, including the dissimilarity index and the diffusion distance. After reviewing the concept of modularity and the basic idea of simulated annealing, we propose our algorithms and the corresponding strategies in Section 3. In Section 4, we apply the proposed methods to the representative examples mentioned before. Finally we give the conclusions in Section 5.

2. The measures of proximity between nodes in networks

2.1. The dissimilarity index and the corresponding center

In Ref. [10], an index of dissimilarity between pairs of nodes is defined, with which one can measure the extent of proximity between nodes of a network. 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 the 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 the node x [17,23]. Suppose the random walker is located at node x. The mean first-passage time $t(x, y)$ is the average number of steps that it takes before it reaches node y for the first time, which is given by

$$t(x, y) = p(x, y) + \sum_{j=1}^{+\infty}(j + 1) \cdot \sum_{z_1,\ldots,z_j\neq y} p(x, z_1)p(z_1, z_2)\cdots p(z_j, y). \tag{2}$$

It has been shown that $t(x, y)$ is the solution of the linear equation

$$[I - B(y)]\begin{pmatrix} t(1, y) \\ \vdots \\ t(n, y) \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{3}$$

where $B(y)$ is the matrix formed by replacing the y-th column of matrix P with a column of zeros [10]. The difference in the perspectives of nodes x and y as regards the network can be quantitatively measured. The dissimilarity index is defined by the following expression:

$$\Lambda(x, y) = \frac{1}{n - 2}\left(\sum_{z\in S, z\neq x,y}\Big(t(x, z) - t(y, z)\Big)^2\right)^{\frac{1}{2}}. \tag{4}$$

We take a partition of S as $S = \bigcup_{k=1}^{N} S_k$ with $S_k \bigcap S_l = \emptyset$ if $k \neq l$. If two nodes x and y belong to the same community, then the average distance $t(x, z)$ will be quite similar to $t(y, z)$; therefore the network's two perspectives will be quite similar. Consequently, $\Lambda(x, y)$ will be small if x and y belong to the same community and large if they belong to different communities. The center $m^I(S_k)$ of community S_k can be defined as

$$m^I(S_k) = \arg\min_{x\in S_k} \frac{1}{|S_k|}\sum_{y\in S_k, y\neq x} \Lambda(x, y), \quad k = 1, \ldots, N. \tag{5}$$

where $|S_k|$ is the number of nodes in community S_k. This is an intuitive idea for choosing the node that reaches others in the same community with the minimal average dissimilarity index as the center.

2.2. The diffusion distance and the geometric centroid

The main idea of Ref. [11] is to define a system of coordinates with an explicit metric that reflects the connectivity of nodes in a given network and the construction is also based on a Markov random walk on networks. This Markov chain has the stationary distribution $\mu(x) = \frac{d(x)}{\sum_{z\in S} d(z)}$ and it satisfies the detailed balance condition $\mu(x)p(x, y) = \mu(y)p(y, x)$. The diffusion distance $D(x, y)$ between x and y is defined as the weighted L^2 distance

$$D^2(x, y) = \sum_{z\in S} \frac{(p(x, z) - p(y, z))^2}{\mu(z)}, \tag{6}$$

where the weight $\mu(z)^{-1}$ penalizes discrepancies on domains of low density more than ones on domains of high density. This notion of proximity of nodes reflects the intrinsic geometry of the set in terms of connectivity of the nodes in a diffusion process. The transition matrix P has a set of left and right eigenvectors and a set of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}| \geq 0$:

$$P\varphi_i = \lambda_i\varphi_i, \qquad \psi_i^T P = \lambda_i\psi_i^T, \quad i = 0, 1, \ldots, n-1. \tag{7}$$

Note that $\psi_0 = \mu$ and $\varphi_0 \equiv 1$. We also have $\psi_i(x) = \varphi_i(x)\mu(x)$. Let q be the largest index i such that $|\lambda_i| > \delta|\lambda_1|$ and if we introduce the diffusion map

$$\Psi : x \longmapsto \begin{pmatrix} \lambda_1\varphi_1(x) \\ \vdots \\ \lambda_q\varphi_q(x) \end{pmatrix}, \tag{8}$$

then the diffusion distance $D(x, y)$ can be approximated to relative precision δ using the first q non-trivial eigenvectors and eigenvalues

$$D^2(x, y) \simeq \sum_{i=1}^{q} \lambda_i^2\Big(\varphi_i(x) - \varphi_i(y)\Big)^2 = \|\Psi(x) - \Psi(y)\|^2. \tag{9}$$

The geometric centroid $c(S_k)$ of community S_k is defined as

$$c(S_k) = \sum_{x\in S_k} \frac{\mu(x)}{\hat{\mu}(S_k)}\Psi(x), \quad k = 1, \ldots, N, \tag{10}$$

where $\hat{\mu}(S_k) = \sum_{x\in S_k} \mu(x)$ [11]. Here $c(S_k)$ may not belong to the set $\{\Psi(x)\}_{x\in S}$. In order to obtain representative centers of the communities that belong to the node set S, we introduce the diffusion center $m^D(S_k)$ via

$$m^D(S_k) = \arg\min_{x\in S_k} \|\Psi(x) - c(S_k)\|^2, \quad k = 1, \ldots, N. \tag{11}$$

3. Simulated annealing to maximize modularity with *k*-means algorithms

In recent years, a concept of modularity proposed by Newman [6–9] has been widely used as a measure of the particular partition of the network into groups, with larger values indicating stronger community structure. For a given partition $\{S_k\}_{k=1}^N$, the modularity Q can be written in the following form:

$$Q = \frac{1}{2m}\sum_{k=1}^{N}\sum_{x,y\in S_k}\Big(e(x, y) - p^E(x, y)\Big), \tag{12}$$

where $p^E(x, y) = \frac{d(x)d(y)}{2m}$. This model has been studied in the past in its own right as a model of a network, and is also closely related to the configuration model, which has been widely studied in the physics literature [8]. Some existing methods are presented for finding good partitions of a network into communities by optimizing the modularity over possible divisions, which has proven highly effective in practice [13]. Our work is different from the earlier ones since we are using simulated annealing to find the maximum of Q with a k-means iterative procedure, which can accelerate the tendency of maximizing the modularity. The algorithms proposed later lead to a more optimal modularity than the existing methods [6–9,12,14,13].

The first simulated annealing algorithm was motivated by simulating the physical process of annealing solids [19]. The process can be described as follows. Firstly, a solid is heated from 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 applied to optimization problems [20].

Let $E = -Q$. $E^{(n)}$ 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 $\exp(-\frac{1}{T}\Delta E^{(n)})$, where $\Delta E^{(n)} = E^{(n+1)} - E^{(n)}$ is the difference of energy and T is the current temperature. Worse solutions are accepted on the basis of the change in solution quality which allows the search to avoid becoming 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 criteria have been met. The initial state is generated at random by N communities, where N is an 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, and 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 R times at the given temperature. A cooling rate $0 < \alpha < 1$ decreased the current temperature until it reached the bound $T_{\min}$. The whole procedure of the simulated annealing with the dissimilarity-index-based k-means algorithm (SADI) is summarized below.

(1) Set parameters $T_{\max}$, $T_{\min}$, $N_{\min}$, $N_{\max}$, 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 centers $\{m^l(S_k^{(0)})\}_{k=1}^{N}$ according to (5), and then calculate the initial energy $E^{(0)}$ using (12); set $n^* = 0$.
(3) For $n = 0, 1, \ldots, R$, do the following:
  (3.1) Generate a set of centers $\{m^l(S_k^{(n)})\}_{k=1}^{N'}$ according to our proposal below and set $N = N'$.
  (3.2) Update the partition $\{S_k^{(n+1)}\}_{k=1}^{N}$ and the center set $\{m^l(S_k^{(n+1)})\}_{k=1}^{N}$ according to

$$S_k^{(n+1)} = \left\{x : k = \arg\min_l \Lambda(x, m^l(S_l^{(n)}))\right\}, \quad k = 1, \ldots, N, \tag{13}$$

  and (5), respectively, then calculate the new energy $E^{(n+1)}$ using (12).
  (3.3) Accept or reject the new state. If $E^{(n+1)} < E^{(n)}$ or $E^{(n+1)} > E^{(n)}$ with $u \sim \mathcal{U}[0, 1]$, $u < \exp\{-\frac{1}{T}\Delta E^{(n)}\}$, then accept the new solution by setting $n = n + 1$; otherwise, reject it.
  (3.4) Update the optimal state, i.e. if $E^{(n)} < E^{(n^*)}$, set $n^* = n$.
(4) The cooling temperature $T = \alpha \cdot T$. If $T < T_{\min}$, go to Step (5); otherwise, set $n = n^*$, and repeat Step (3).
(5) Output the optimal solution $\{S_k^{(n^*)}\}_{k=1}^{N}$ and the minimum energy $E^{(n^*)}$ of the whole procedure.

We can also obtain the simulated annealing with the diffusion-distance-based k-means algorithm (SADD) by replacing (5) and (13) with (11) and

$$S_k^{(n+1)} = \left\{x : k = \arg\min_l \|\Psi(x) - c(S_l^{(n)})\|^2\right\}, \quad k = 1, \ldots, N, \tag{14}$$

in the above procedure. Our proposal for the process of generating a set of new centers in Step (3.1) comprises three functions, including retaining a current community, deleting a current community and splitting a current community. At each iteration, one of the three functions can be randomly chosen and the community strength [24]

$$M(S_k) = \sum_{x \in S_k} (d^{\text{in}}(x) - d^{\text{out}}(x)), \quad k = 1, \ldots, N, \tag{15}$$

is used to select a community, where $d^{\text{in}} = \sum_{z \in S_k} e(x, z)$ and $d^{\text{out}} = \sum_{z \notin S_k} e(x, z)$. The three functions are described below

- Retain Community. We retain the center set.
- Delete Community. The community with the minimum community strength S_d is identified using (15) and its center should be deleted from the center set.
- Split Community. The community with the minimum average community strength

$$S_s = \arg\min_{S_l} \frac{M(S_l)}{|S_l|} \tag{16}$$

  is chosen. For SADI, the new center is obtained by using

$$m^l(S_{N+1}) = \arg \min_{x \in S_s, x \neq m(S_s)} \Lambda(x, m^l(S_s)), \tag{17}$$

  and for SADD, the current $c(S_s)$ is replaced by two new geometric centers created by

$$\begin{aligned} c(S_{N+1}) &= c(S_s) - |c(S_s) - m(S_s)|, \\ c(S_s) &= c(S_s) + |c(S_s) - m(S_s)|. \end{aligned} \tag{18}$$

The number of the iteration steps depends on the initial and terminal temperature, the cooling rate and the repeating times at the given temperature, which is about $R \log_\alpha \frac{T_{\min}}{T_{\max}}$. The bounds for N are usually chosen as $N_{\min} = 2$ and $N_{\max} = n/3$.

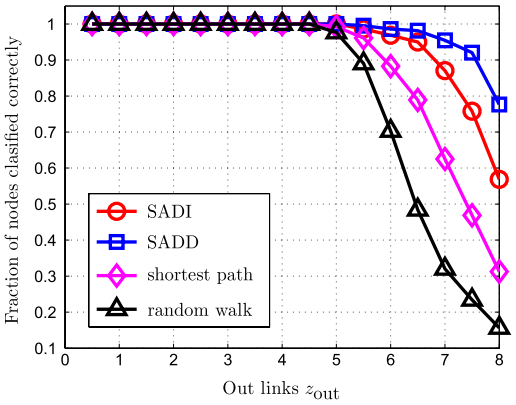

Fig. 1. The fraction of nodes classified correctly for the ad hoc network obtained by the proposed algorithms, as compared with the existing methods of Ref. [6]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For SADI, since what we really need to know is the difference of mean first-passage times, i.e. $t(x, z)-t(y, z)$, we can calculate all the different differences with a computational time of $O(n^3)$. For SADD, the computation of both $\{c(S_k)\}_{k=1}^N$ and $\{S_k\}_{k=1}^N$ costs $O(qNn)$ for each iteration.

The advantage of our algorithms is that they overcome the weaknesses of the traditional clustering search that the optimal prediction error is decreasing as the number of the clusters increases, such as in the k-means and fuzzy c-means algorithms constructed from the gradient descent method [13,15,16,18]. The cooling process of our methods can efficiently and automatically determine the number of communities N without fixing it as a known model parameter, and the initial partition $\{S_k^{(0)}\}_{k=1}^N$ can be randomly chosen. The problem can be solved by means of another approach of searching over all possible N using the two k-means algorithms. But this will have extremely high cost since for each fixed N, the k-means procedure should be operated for 1000 to 5000 trials due to its local minimum. On the other hand, our methods can sometimes obtain a more optimal partitioning result with a larger value of modularity than searching over all possible N using k-means (see Fig. 4). Hence, the simulated annealing strategy can avoid ineffectively repetition and lead to a high degree of efficiency and accuracy.

4. Experimental results

4.1. Ad hoc networks 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 [5,6,10,12,13,15,16]. It has a known partition and is constructed as follows. Suppose we choose $n = 128$ nodes, and split them into 4 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 such that the average node degree d is fixed at $d = 16$. In other words, p_{in} and p_{out} are related as

$$31p_{\text{in}} + 96p_{\text{out}} = 16. \tag{19}$$

We will define $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. We change z_{out} from 0.5 to 8 and look into the fraction of nodes which are correctly classified. The parameters are set as $T_{\max} = 3.0$, $T_{\min} = 0.01$, $\alpha = 0.9$ and $R = 20$ in this model computation. The fraction of correctly identified nodes is shown in Fig. 1, comparing against the two methods described in Ref. [6]. It seems that SADI and SADD perform noticeably better than the two previous methods, especially for the more difficult cases when z_{out} is large.

4.2. A sample network generated from a Gaussian mixture model

To further test the validity of the algorithms, we apply them to a sample network generated from a Gaussian mixture model. This model is quite closely related to the random geometric graph except that we take a Gaussian mixture here compared with the uniform distribution in Ref. [25].

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

$$\sum_{k=1}^{K} q_k G\left(\boldsymbol{\mu}_k, \Sigma_k\right), \tag{20}$$

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

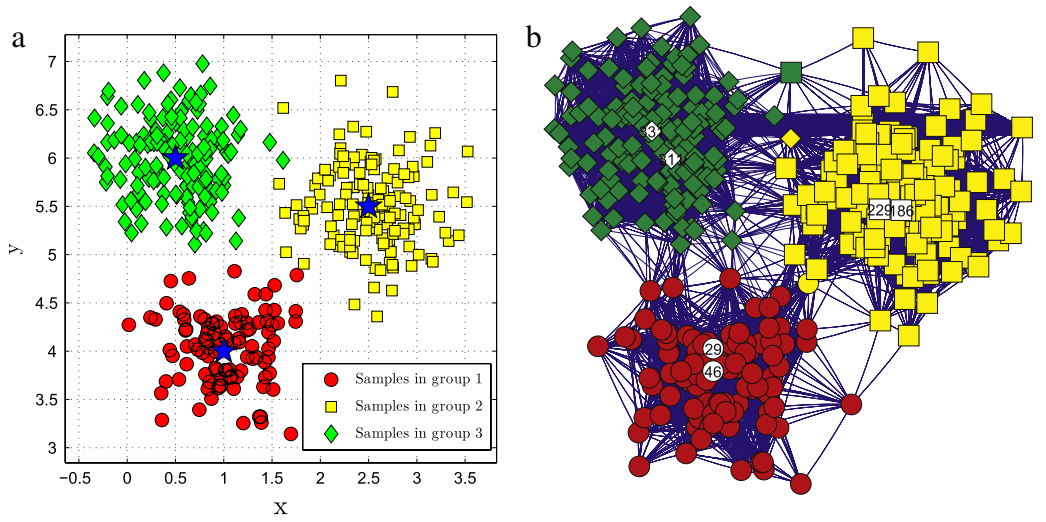

Fig. 2. (a) 400 sample points generated from the given three-Gaussian mixture distribution. The star symbols represent the centers of each Gaussian component. The circle, square and diamond shaped symbols represent the positions of sample points in each component. (b) The partition for the network generated from the sample points in (a) with dist $= 0.8$. The communities are represented by different colors. The centers $m^I = \{46, 186, 331\}$ and $m^D = \{29, 229, 311\}$ are in white.

Table 1
The numerical results obtained by our proposed methods for the karate club network, the dolphins network and the political books network.

	The karate club network		The dolphins network		The political books network	
	N	Q	N	Q	N	Q
SADI	4	0.4198	5	0.5176	4	0.5260
SADD	4	0.4174	4	0.5235	4	0.5266

is, if $|\boldsymbol{x}_i - \boldsymbol{x}_j| \leq$ 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 the metric. We take $n = 400$ and $K = 3$, and then generate the sample points with the means and the covariance matrices

$$\boldsymbol{\mu}_1 = (1.0, 4.0)^T, \qquad \boldsymbol{\mu}_2 = (2.5, 5.5)^T, \qquad \boldsymbol{\mu}_3 = (0.5, 6.0)^T, \tag{21a}$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{21b}$$

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, approximately $q_1 = 100/400$, $q_2 = q_3 = 150/400$. We take dist $= 0.8$ in this example. The sample points are shown in Fig. 2(a) and the partition result is shown in Fig. 2(b). Here we implement the algorithms by setting $T_{\max} = 3.0$, $T_{\min} = 0.01$, $\alpha = 0.9$ and $R = 15$. SADI and SADD produce the same result with $N = 3$ and $Q = 0.6241$, and the centers are given by $m^I = \{46, 186, 331\}$ and $m^D = \{29, 229, 311\}$. The mean L^2-error between the centers obtained and the means $\boldsymbol{\mu}$

$$\frac{1}{N}\sum_{k=1}^{N} \|\boldsymbol{x}_{m(S_k)} - \boldsymbol{\mu}_k\|_2 \tag{22}$$

for the two methods are 0.0804 and 0.2211 respectively. The results are extremely reasonable, which indicates that our algorithms go smoothly with several hundreds of nodes.

4.3. Some real-world networks

In this subsection, we describe the applications of the algorithms to five further real-world networks, including the karate club network, the dolphins network, the political books network, the network of characters in the novel *Les Misérables* and the American football team network.

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 [26]. Soon after, a dispute arose between the club administrator and the main teacher and the club split into two smaller clubs. It has been used widely to test algorithms for finding communities in networks [5–10,12,14–16]. The results obtained by our methods are shown in Table 1 and Fig. 3. Both methods obtain

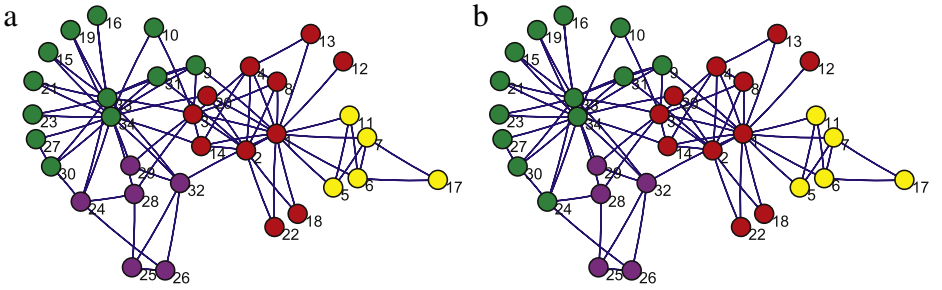

Fig. 3. The community structure of the karate club network detected by the proposed methods. (a) The result detected by SADI. (b) The result detected by SADD. The two methods produce nearly the same outcome, except for node 24.

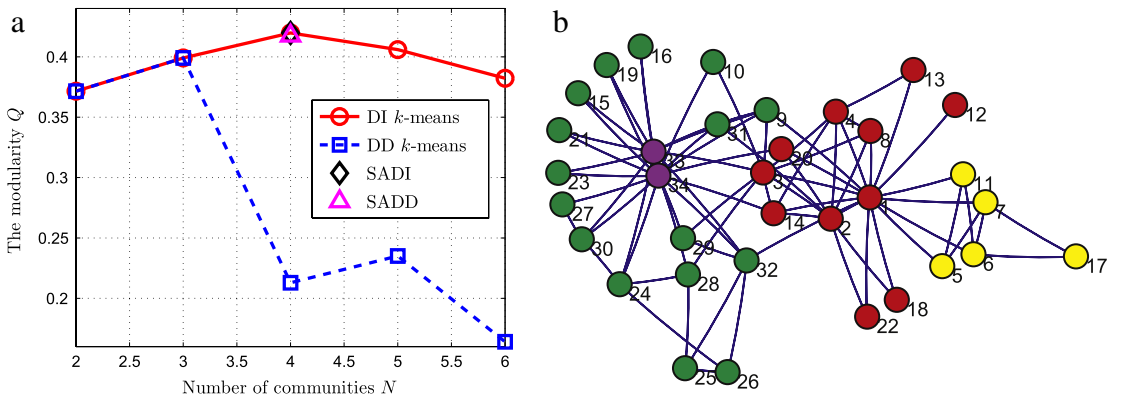

Fig. 4. (a) The greatest modularity detected using dissimilarity-index-based and diffusion-distance-based k-means, SADI and SADD. It shows clearly that SADI can find the maximal value of $Q = 0.4198$ obtained by searching over N using the corresponding k-means while SADD can reach a larger modularity at $Q = 0.4174$ than the corresponding k-means. (b) The community structure obtained using diffusion-distance-based k-means with $N = 4$ and in the dissimilarity-index-based case is the same as that of Fig. 3(a).

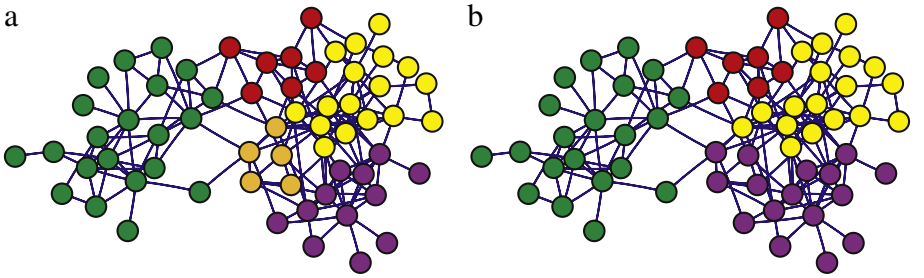

Fig. 5. The community structure of the dolphins network detected by the proposed methods. (a) The result detected by SADI. (b) The result detected by SADD.

$N = 4$, while SADI reaches a higher value of Q and outperforms most of the existing methods [6,12,14]. On the other hand, the simulated annealing approach can obtain a more optimal partitioning result than searching over all possible N using the two k-means algorithms, which are illustrated in detail in Fig. 4.

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 [27]. The network was compiled from studies of the dolphins, with ties between dolphin pairs being established by the observation of statistically significant frequent association [6]. The results obtained by our methods are shown in Table 1 and Fig. 5. According to the results, the network seems to be split into two large communities: the green part and the larger one; and the larger one keeps splitting into a few smaller communities, represented by different colors. The split into two groups appears to correspond to a known division of the dolphin community [28]. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals, in that the yellow part consists almost of entirely of females and the others almost entirely of males.

The political books network. We consider the network of books on politics, which are assigned on the basis of a reading of the descriptions and reviews of the books posted on Amazon [9]. 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

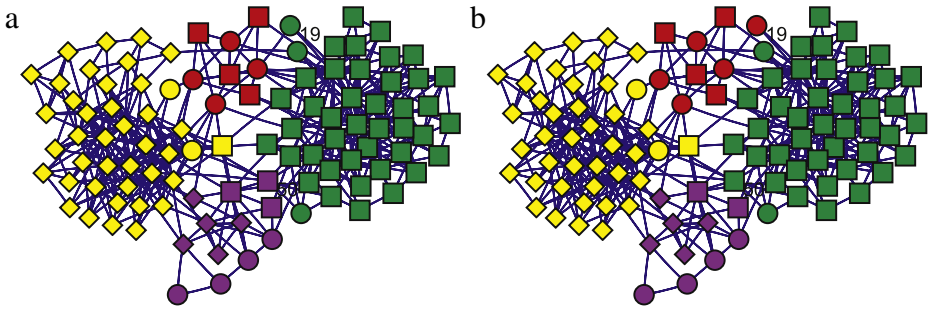

Fig. 6. The community structure of the political books network detected by the proposed methods. (a) The result detected by SADI. (b) The result detected by SADD. The two methods produce nearly the same outcome except for nodes 19 and 50. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

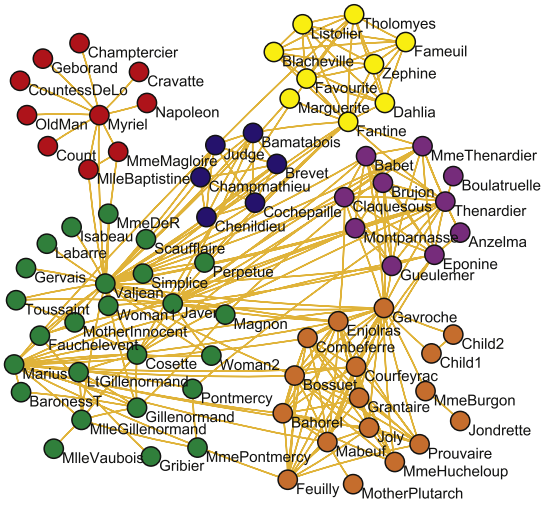

Fig. 7. The community structure of the network of interactions between major characters in the novel *Les Misérables* by Victor Hugo. The greatest modularity achieved by using SADD is $Q = 0.5654$ and corresponds to the six communities represented by the colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

purchased by the same buyer, as indicated by the feature that customers who bought this book also bought the other books. As shown in Fig. 6, nodes have been given according to whether they are conservative (box) or liberal (diamond), except for a small number of books which are neutral (ellipse). The results are shown in Table 1 and Fig. 6. We find four 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 groups. Thus these books appear to form communities for copurchasing that align closely with political views.

The network of characters in the novel Les Misérables. This is the network of interactions between major characters in Victor Hugo's sprawling novel of crime and redemption in post-restoration France, *Les Misérables* [6]. Using the list of character appearances by scene, a network was constructed in which the nodes represent characters and the edge between two nodes represents co-appearance of the corresponding characters in one or more scenes. The optimal community structure of the result obtained by SADD has a modularity of $Q = 0.5654$ and gives six communities as shown in Fig. 7. It can obtain a larger modularity than the methods in Ref. [6]. The communities clearly reflect the subplot structure of the book: the protagonist Jean Valjean and his nemesis, the police officer Javert, are central to the network and form the hubs of communities composed of their respective adherents. Other subplots centered on Marius, Cosette, Fantine and the bishop Myrial are also shown in Fig. 7.

The football team network. The last network that we investigated is the college football network which represents the game schedule of the 2000 season of Division I of the US college football league [5]. The nodes in the network represent 115 teams and edges represent regular season games between the two teams that they connect. The teams are divided into conferences containing around 8 to 12 each. Games are more frequent between members of the same conference than between members of different conferences. The optimal community structure split of the network obtained using SADI has a strong modularity of $Q = 0.6032$ and gives 11 communities as shown in Fig. 8, outperforming most existing methods [5,10,14]. According to the results, 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 seem

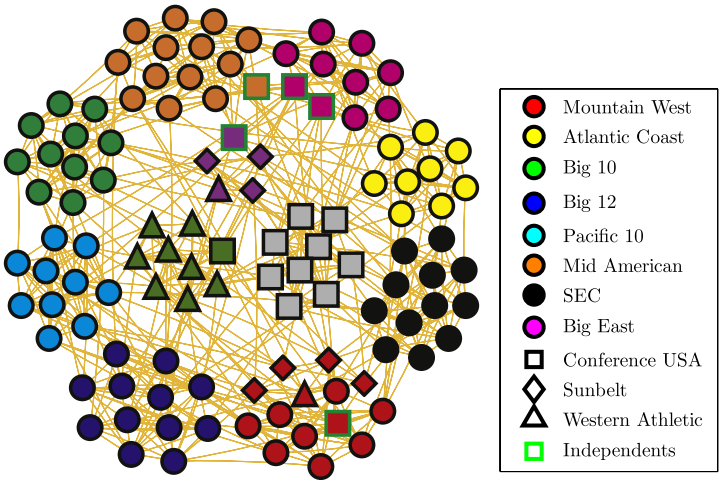

Fig. 8. The community structure of the American football team network. The greatest modularity achieved by using SADI is $Q = 0.6032$ and it corresponds to the 11 communities represented by the colors.

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 is split into two communities; one is clustered with a team less connected with the Western Athletic conference and the other is clustered with Mountain West. Only one member in Conference USA is grouped with most of the teams in the Western Athletic conference. All the other communities coincide with the known structure.

5. Conclusions

In this paper, we present a new method for detecting the community structure in complex networks. The proposed algorithms, simulated annealing using dissimilarity-index-based k-means (SADI) and simulated annealing using diffusion-distance-based k-means (SADD), are constructed and successfully applied to several representative networks. The numerical results show that they produce similar results, but the SADD has better efficiency and accuracy in most cases. Both of the algorithms outperform the existing methods in the literature [5–10,12,14] as regards optimal modularity. We again point out that our algorithms can not only find the community structure, but also identify the central node of each community. The optimal number of communities can be efficiently determined without any prior knowledge about the community structure during the cooling process.

The optimal prediction error decreases as the number of clusters increases in the k-means framework. This means that the model selection strategy is not self-contained with this proposal. Investigating how to draw a lesson from the validity index in the traditional clustering literature [29] for complex networks utilizing the corresponding optimal prediction error instead of modularity to obtain a better quantity in evaluating the partition will be our next step. But the algorithms considered in this paper are efficient and deserved to be investigated.

Acknowledgements

This work was supported by the Natural Science Foundation of China under grant 10871010 and the National Basic Research Program of China under grant 2005CB321704. The authors thank Professor M.E.J. Newman for providing the data for the karate club network, the dolphins network, the political books network, the network of characters in the novel *Les Misérables* and the American football team network.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97.
[2] M. Newman, A.-L. Barabási, D.J. Watts, The Structure and Dynamics of Networks, Princeton University Press, Princeton, 2005.
[3] 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.
[4] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, A. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[5] M. Girvan, M. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.
[6] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[7] M. Newman, Detecting community structure in networks, Eur. Phys. J. B 38 (2) (2004) 321–330.
[8] M. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[9] M. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103 (23) (2006) 8577–8582.
[10] H. Zhou, Distance, dissimilarity index, and network community structure, Phys. Rev. E 67 (6) (2003) 061901.

[11] 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.
[12] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[13] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech. 9 (2005) P09008.
[14] J. Zhang, S. Zhang, X. Zhang, Detecting community structure in complex networks based on a measure of information discrepancy, Physica A 387 (2008) 1675–1682.
[15] 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.
[16] T. Li, J. Liu, W. E, Probabilistic framework for network partition, Phys. Rev. E 80 (2009) 026106.
[17] L. Lovasz, Random walks on graphs: A survey, Combinatorics, Paul Erdos is Eighty 2 (1993) 1–46.
[18] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
[19] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087.
[20] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[21] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.
[22] V. Latora, M. Marchiori, A measure of centrality based on the network efficiency, New J. Phys. 9 (2007) 188.
[23] F. Chung, Spectral Graph Theory, American Mathematical Society, Rhode Island, 1997.
[24] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101 (9) (2004) 2658–2663.
[25] M. Penrose, Random Geometric Graphs, Oxford University Press, Oxford, 2003.
[26] W. Zachary, An information flow model for conflict and fission in small groups, J. Anthrop. Res. 33 (4) (1977) 452–473.
[27] D. Lusseau, The emergent properties of a dolphin social network, Proc. Roy. Soc. B: Biol. Sci. 270 (2003) 186–188.
[28] 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.
[29] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 841–847.