An Extended Validity Index for Identifying Community Structure in Networks

Jian Liu

LMAM and School of Mathematical Sciences, Peking University,
Beijing 100871, P.R. China
dugujian@pku.edu.cn

Abstract. To find the best partition of a large and complex network into a small number of communities has been addressed in many different ways. In this paper, a new validity index for network partition is proposed, which is motivated by the construction of Xie-Beni index in Euclidean space. The simulated annealing strategy is used to minimize this extended validity index, associating with a dissimilarity-index-based k-means iterative procedure, under the framework of a random walker Markovian dynamics on the network. The proposed algorithm(SAEVI) can efficiently and automatically identify the community structure of the network and determine an appropriate number of communities without any prior knowledge about the community structure during the cooling process. The computational results on several artificial and real-world networks confirm the capability of the algorithm.

Keywords: Validity index, Community structure, Dissimilarity index, Simulated annealing, K-means.

1 Introduction

In recent years we have seen an explosive growth of interest and activity on 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, powergrid, etc, due to our increased capability of analyzing these models. 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 clusters [3,4,5,6,7,8], which are constructed from different viewing angles comparing different proposals in the literature. 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 [9,10].

L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISNN 2010, Part II, LNCS 6064, pp. 258–267, 2010.
© Springer-Verlag Berlin Heidelberg 2010

In [6], a dissimilarity index for each pair of nodes and the corresponding hierarchical algorithm to partition the networks are proposed. The basic idea is to associate the network with the random walker Markovian dynamics [11]. This can motivate us to solve the partition problem by an analogy to the traditional k-means algorithm [12] based on this dissimilarity index. In traditional clustering literature, a function called validity index [13] is often used to evaluate the quality of clustering results. The optimal number of clusters can be determined by selecting the minimal value of the index. We construct an extended formulation of Xie-Beni index [13], which has smaller values indicating stronger community structure in networks. Then simulated annealing strategy [14,15] is utilized to obtain the minimal value of the index, associating with a dissimilarity-index-based k-means procedure.

We will construct our algorithm — simulated annealing to minimize the extended validity index (SAEVI) for network partition. From the numerical performance to four model problems: the ad hoc network with 128 nodes, sample networks generated from Gaussian mixture model, the karate club network and the American football team network, we can see that our algorithm can efficiently and automatically determine the optimal number of communities and identify the community structure during the cooling process.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the dissimilarity index [6] which signifies to what extent two nodes would like to be in the same community, then proposed the extended validity index for network partition. After reviewing the idea of simulated annealing, we describe our algorithm (SAEVI) and the corresponding strategies in Section 3. In Section 4, we apply the algorithm to four representative examples mentioned before. Finally we make the conclusion in Section 5.

2 The Framework for Network Partition

2.1 The Dissimilarity Index and the Corresponding Center

In [6], a dissimilarity index between pairs of nodes is defined, 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 nodes 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 [7,8,11]. Suppose the random walker is located at node x. The mean first passage time $t(x,y)$ is the average number of steps it takes before it reaches node y for the first time, which is given by

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

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 [6]. The difference in the perspectives of nodes x and y about 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 $m(S_k)$ of community S_k can be defined as

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

where $|S_k|$ is the number of nodes in community S_k. This is an intuitive and reasonable idea for us to choose the node reached others in the same community with the minimum average dissimilarity index as the center of S_k.

2.2 The Extended Xie-Beni Index

A well known validity index for fuzzy clustering called Xie-Beni index [13] is widely used to classify samples overlap in Euclidean space, which is based on the fuzzy c-means algorithm [12]. The basic idea of FCM algorithm is to minimize the following objective function with respect to the fuzzy memberships $\rho_k(x)$ and the cluster centers m_k

$$J(\rho, m) = \sum_{k=1}^{N} \sum_{x \in S} \rho_k^b(x) \|x - m_k\|^2, \quad b \geq 1, \tag{6}$$

where $b > 1$ is the fuzziness index. For the FCM algorithm with $b = 2$, Xie-Beni index V_{XB} can be explicitly written as

$$V_{XB} = \frac{\sum_{x \in S} \sum_{k=1}^{N} \rho_k^2(x) \|x - m_k\|^2}{n \min_{k \neq l} \|m_k - m_l\|^2} = \frac{J(\rho, m)}{nK(m)}. \tag{7}$$

Here $J(\rho, m)$ measures the compactness of the data set S and $K(m)$ measures the separation. The more separate the clusters, the larger $K(m)$ and the smaller V_{XB}. We can find an optimal cluster number by solving $\min_{2 \leq N \leq n-1} V_{XB}$ to produce a best clustering performance for the data set S.

We extend the idea of considering both compactness and separateness to our formulation, and propose a new validity index for network partition as following

$$V_E = \frac{\sum_{k=1}^{N} \sum_{x \in S_k} \Lambda^2(x, m_k)}{\min_{k \neq l} \Lambda^2(m(S_k), m(S_l))} = \frac{J_E}{K_E}. \qquad (8)$$

where J_E is the objective function constructed for the dissimilarity-index-based k-means which reflects compactness, and K_E plays the role of separation such as $K(m)$ in (7). An ideal partition requires a more stable state in space $\mathbb{S} = \{S_1, \ldots, S_N\}$, which has smaller J_E and larger K_E. Thus, an optimal partition can be find by solving

$$\min_{N} \left\{ \min_{\{S_1, \cdots, S_N\}} V_E \right\}. \qquad (9)$$

The global optimal problem (9) can be solved by searching over the all possible N with k-means algorithm. But this will cost extremely much, since for each fixed N, the k-means procedure should be operated 1000 to 5000 trials due to its local minima. However, the simulated annealing strategy [14,15] can avoid repeating ineffectively and lead to a high degree of efficiency and accuracy.

3 The Algorithm

The simulated annealing strategy is utilized here to address (9), which is motivated by simulating the physical process of annealing solids [14]. 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 used to optimization problems [15].

Let $E = V_E$. $E^{(n)}$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 is generated by randomly N clusters, here $N \in [N_{\min}, N_{\max}]$, and the initial temperature T is set to a high temperature $T_{\max}$. A neighbor of the current state is produced by randomly flipping one spin, 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 minimize the Extended Validity Index (SAEVI) with k-means algorithm is summarized as follows

(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 partition $\{S_k^{(0)}\}_{k=1}^N$ randomly; Set the current temperature $T = T_{\max}$.

(2) Compute the centers $\{m(S_k^{(0)})\}_{k=1}^N$ according to (5), then calculate the initial energy $E^{(0)}$ using the definition of V_E (8); Set $n^* = 0$.

(3) For $n = 0, 1, \cdots, R$, do the following

(3.1) Generate a set of centers $\{m(S_k^{(n)})\}_{k=1}^{N'}$ according to our proposal below and set $N = N'$;

(3.2) Update the partition $\{S_k^{(n+1)}\}_{i=1}^N$ using

$$S_k^{(n+1)} = \left\{x : k = \arg\min_l \varLambda(x, m(S_l^{(n)})\right\}, \quad k = 1, \cdots, N, \tag{10}$$

and the corresponding $\{m(S_k^{(n+1)})\}_{k=1}^N$ according to (5), then calculate the new energy $E^{(n+1)}$ using (8);

(3.3) Accept or reject the new state. If $E^{(n+1)} < E^{(n)}$ or $E^{(n+1)} > E^{(n)}$ with $u \sim \mathcal{U}[0,1]$, $u < \exp\{-\frac{1}{T}\triangle E^{(n)}\}$, then accept the new solution by setting $n = n + 1$; Else,reject it;

(3.4) Update the optimal state, i.e. 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 $\{S_k^{(n^*)}\}_{k=1}^N$ and the minimum energy $E^{(n^*)}$ of the whole procedure.

Our proposal to the process of generating a new partition 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 strength [16]

$$M(S_k) = \sum_{x \in S_k} (d^{\text{in}}(x) - d^{\text{out}}(x)), \quad k = 1, \cdots, N, \tag{11}$$

is used to select a center, 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

(i) Delete Community. The community with the minimal community strength S_d is identified using (11) and its center should be deleted from $\{m(S_k)\}_{k=1}^N$.

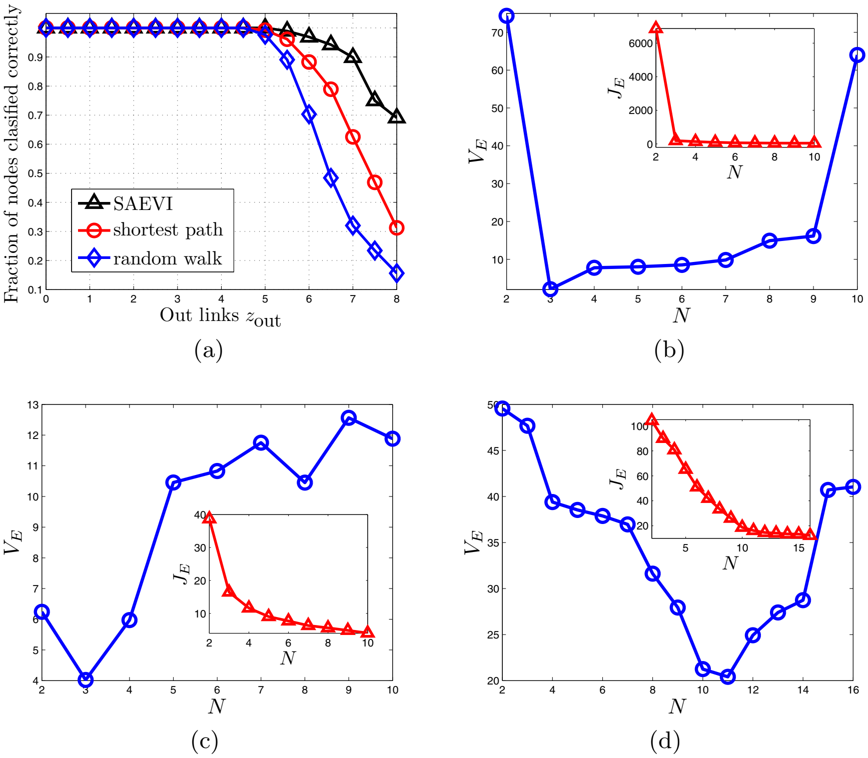
(ii) Split Community. The community with the minimal average community strength

$$S_s = \arg\min_{S_l} \frac{M(S_l)}{|S_l|} \tag{12}$$

is chosen. The new center is obtained by

$$m(S_{N+1}) = \arg\min_{x \in S_s, x \neq m(S_s)} \varLambda(x, m(S_s)). \tag{13}$$

(iii) Keep Community. We maintain the center set $\{m(S_k)\}_{k=1}^N$.



(a)

(b)

(c)

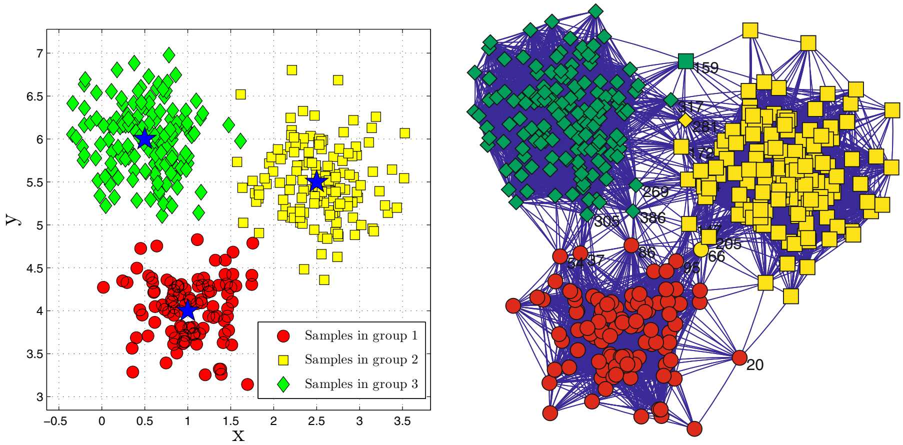
(d)

Fig. 1. (a)The fraction of nodes classified correctly of ad hoc network by SAEVI compared with the methods in [4]. (b)V_E and J_E changed with N for the network generated from the given 3-Gaussian mixture model. (c)V_E and J_E changed with N for the karate club network. (d)V_E and J_E changed with N for the football team network.

4 Experimental Results

4.1 Ad Hoc Network with 128 Nodes

We apply our methods to the ad hoc network with 128 nodes. The ad hoc network is a typical benchmark problem considered in many papers [4,6,7,8]. Suppose we choose $n = 128$ nodes, split into 4 communities containing 32 nodes each. Assume 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, d, is fixed at $d = 16$. In other words p_{in} and p_{out} are related as $31p_{\text{in}}+96p_{\text{out}} = 16$. Here we naturally choose the nodes group $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. We change z_{out} from 0.5 to 8 and look into the fraction of nodes which correctly classified. The fraction of correctly identified nodes is shown in Figure 1(a), comparing with the two methods described in [4]. It seems that SAEVI performs noticeably



(a) (b)

Fig. 2. (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 shaped symbols represent the position of sample points in each component, respectively; (b)The partition for the network generated from the sample points in (a) with $dist = 0.8$. The optimal extended validity index achieved is $V_E = 2.1130$ and corresponds to the 3 communities represented by the colors.

better than the two previous methods, especially for the more diffusive cases when z_{out} is large.

4.2 Sample Network Generated from 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 quite related the concept random geometric graph proposed by Penrose [17]. We generate n sample points $\{\boldsymbol{x}_i\}$ in two dimensional Euclidean space subject to a K-Gaussian mixture distribution $\sum_{k=1}^{K} q_k G\left(\boldsymbol{\mu}_k, \varSigma_k\right)$, where $\{q_k\}$ are mixture proportions satisfying $0 < q_k < 1$, $\sum_{k=1}^{K} q_k = 1$. $\boldsymbol{\mu}_k$ and $\varSigma_k$ are the mean positions and covariance matrices for each component, respectively. Then we generate the network as following: 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. We take $n = 400$ and $K = 3$, then generate the sample points with the means and the covariance matrices as follows

$$\boldsymbol{\mu}_1 = (1.0, 4.0)^T, \boldsymbol{\mu}_2 = (2.5, 5.5)^T, \boldsymbol{\mu}_3 = (0.5, 6.0)^T, \tag{14a}$$

$$\varSigma_1 = \varSigma_2 = \varSigma_3 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{14b}$$

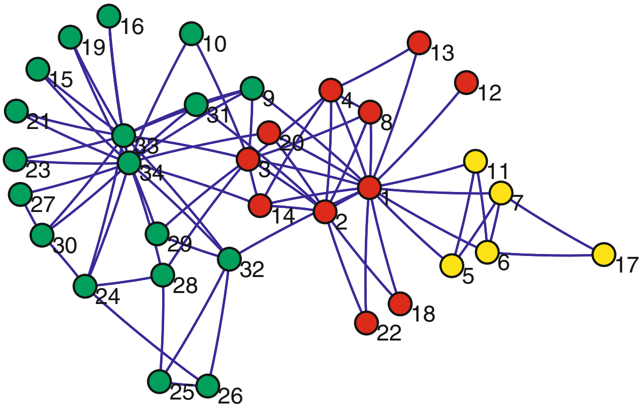


Fig. 3. The partition for the karate club network. The optimal extended validity index achieved is $V_E = 4.0225$ and corresponds to the 3 communities represented by the colors.

Here we pick nodes 1:100 in group 1, nodes 101:250 in group 2 and nodes 251:400 in group 3 for simplicity (see Figure 2(a)). With this choice, approximately $q_1 = 100/400, q_2 = q_3 = 150/400$. The thresholding is chosen as $dist = 0.8$. The numerical and partitioning results obtained by SAEVI are shown in Figure 1(b) and Figure 2(b). The objective function of k-means J_E is decreasing as N increases, while the extended validity index V_E has a minimum.

4.3 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 [18]. Soon after, a dispute arose between the clubs administrator and main teacher and the club split into two smaller clubs. It has been used in several papers to test the algorithms for finding community structure in networks [3,4,5,6,7,8]. The numerical and partitioning results obtained by SAEVI are shown in Figure 1(c) and Figure 3, which seem consistent with the original structure of the network.

4.4 The Football Team Network

The last network we investigated is the college football network which represents the game schedule of the 2000 season of Division I of the US college football league [3,6]. The nodes in the network represent 115 teams and edges represent regular season games between the two teams they connect. The teams are divided into conferences containing around 8 to 12 each. Games are more frequent between members of the same conference than between members of different conferences. The numerical and partitioning results obtained by SAEVI are shown in Figure 1(d) and Figure 4. According to the results, almost all of the football

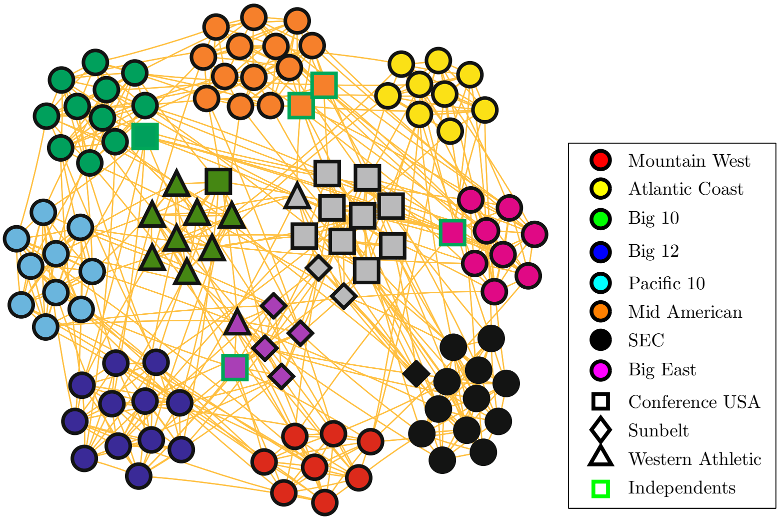


Fig. 4. The partition for the American football team network. The optimal extended validity index achieved is $V_E = 20.4117$ and corresponds to the 11 communities represented by the colors.

teams are correctly clustered with the others in their conference. The teams in Independents conference seem not belonging 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 three communities, clustered with a team which is less connected in Western Athletic conference, Conference USA and SEC. Only one member in Conference USA is grouped with most of the teams in Western Athletic conference. All the other communities coincide with the known structure and our algorithm performs remarkably well.

5 Conclusions

In this paper, we have proposed a new validity index for network partition and used the simulated annealing strategy to minimize this index associating with a dissimilarity-index-based k-means procedure. The algorithm (SAEVI) is constructed and succeeds in four representative examples. It is demonstrated by experiments that our algorithm can identify the community structure with a high degree of accuracy. The optimal number of communities can be efficiently determined without any prior knowledge about the community structure during the cooling process. The proposed validity index is competitive with the modularity for network community structure in the literature [4,5] which leads to the model selection problem. However, the new validity index and the algorithm considered in this paper are efficient and deserve to be investigated.

Acknowledgements. 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. The author thanks Professor M.E.J. Newman for providing the data of the karate club network and the college football team network.

References

1. Albert, R., Barabási, A.L.: Statistical Mechanics of Complex Networks. Rev. Mod. Phys. 74(1), 47–97 (2002)
2. Newman, M., Barabási, A.L., Watts, D.J.: The Structure and Dynamics of Networks. Princeton University Press, Princeton (2005)
3. Girvan, M., Newman, M.: Community Structure in Social and Biological Networks. Proc. Natl. Acad. Sci. USA 99(12), 7821–7826 (2002)
4. Newman, M., Girvan, M.: Finding and Evaluating Community Structure in Networks. Phys. Rev. E 69(2), 026113 (2004)
5. Newman, M.: Modularity and Community Structure in Networks. Proc. Natl. Acad. Sci. USA 103(23), 8577–8582 (2006)
6. Zhou, H.: Distance, Dissimilarity Index, and Network Community Structure. Phys. Rev. E 67(6), 061901 (2003)
7. Weinan, E., Li, T., Vanden-Eijnden, E.: Optimal Partition and Effective Dynamics of Complex Networks. Proc. Natl. Acad. Sci. USA 105(23), 7907–7912 (2008)
8. Li, T., Liu, J., Weinan, E.: Probabilistic Framework for Network Partition. Phys. Rev. E 80, 026106 (2009)
9. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intel. 22(8), 888–905 (2000)
10. Meilă, M., Shi, J.: A Random Walks View of Spectral Segmentation. In: Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, pp. 92–97 (2001)
11. Lovasz, L.: Random Walks on Graphs: A Survey. Combinatorics, Paul Erdos is Eighty 2, 1–46 (1993)
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2001)
13. Xie, X.L., Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Tran. Pattern Anal. Mach. Intel. 13(8), 841–847 (1991)
14. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. 21(6), 1087 (1953)
15. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science 220(4598), 671–680 (1983)
16. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and Identifying Communities in Networks. Proc. Natl. Acad. Sci. USA 101(9), 2658–2663 (2004)
17. Penrose, M.: Random Geometric Graphs. Oxford University Press, Oxford (2003)
18. Zachary, W.: An Information Flow Model for Conflict and Fission in Small Groups. J. Anthrop. Res. 33(4), 452–473 (1977)